M. G. KENDALL, Sc.D.

A COURSE IN MULTIVARIATE ANALYSIS

BEING
NUMBER TWO
OF

GRIFFIN'S STATISTICAL
MONOGRAPHS & COURSES
EDITED BY M. G. KENDALL, Sc.D.

PUBLISHERS' NOTE

The series in which this title appears was introduced by the publishers in 1957 and is under the general editorship of Maurice G. Kendall, Sc.D., Professor of Statistics in the University of London. It is intended to fill a need which has been evident for some time and is likely to grow: the need for some form of publication at moderate cost which will make accessible to a group of readers specialized studies in statistics or special courses on particular statistical topics. There are numerous cases where, for example, a monograph on some newly developed field would be very useful, but the subject has not reached the stage where a comprehensive book is possible; or again, where a course of study is desired in a domain not covered by text-books but where an exhaustive treatment, even if possible, would be expensive and perhaps too elaborate for the readers' needs.

Considerable attention has been given to the problem of producing these books speedily and economically. Appearing in a cover the design of which will be standard, the contents of each volume will follow a simple, straightforward layout, the text production method adopted being suited to the complexity or otherwise of the subject.

The publishers will be interested in approaches from any authors who have work of importance suitable for the series.

CHARLES GRIFFIN & CO. LTD.

OTHER GRIFFIN BOOKS ON STATISTICS, &c.

A statistical primer	F. N. DAVID
An introduction to the theory of statistics	G. U. YULE and M. G. KENDALL
The advanced theory of statistics (three volumes)	M. G. KENDALL and A. STUART
Rank correlation methods	M. G. KENDALL
Exercises in theoretical statistics	M. G. KENDALL
Rapid statistical calculations	M. H. QUENOUILLE
The design and analysis of experiment	M. H. QUENOUILLE
Sampling methods for censuses and surveys	F. YATES
Statistical method in biological assay	D. J. FINNEY
Biomathematics	C. A. B. SMITH
The mathematical theory of epidemics	N. T. J. BAILEY
Probability and the weighing of evidence	I. J. GOOD

GRIFFIN'S STATISTICAL MONOGRAPHS AND COURSES:

No. 1: *The analysis of multiple time-series*	M. H. QUENOUILLE
No. 2: *A course in multivariate analysis*	M. G. KENDALL
No. 3: *The fundamentals of statistical reasoning*	M. H. QUENOUILLE
No. 4: *Basic ideas of scientific sampling*	A. STUART
No. 5: *Characteristic functions*	E. LUKACS
No. 6: *An introduction to infinitely many variates*	E. A. ROBINSON
No. 7: *Mathematical methods in the theory of queueing*	A. Y. KHINTCHINE
No. 8: *A course in the geometry of n dimensions*	M. G. KENDALL

A COURSE IN MULTIVARIATE ANALYSIS

M. G. KENDALL, Sc.D.

Professor of Statistics in the University of London

President, Royal Statistical Society, 1961–62

BEING NUMBER TWO OF

GRIFFIN'S STATISTICAL MONOGRAPHS & COURSES

EDITED BY

M. G. KENDALL, Sc.D.

Second impression

CHARLES GRIFFIN & COMPANY LIMITED

LONDON

First published 1957
Second impression (with minor corrections) ... 1961

Made and printed in Great Britain
Varityped by J. R. Hawkins (London) Ltd, Erith, Kent
Photolithographed by Chorley & Pickersgill Ltd, Leeds

PREFACE

This course was given at the Institute of Statistics of the Consolidated University of North Carolina in the spring of 1954: again at Blacksburg, Virginia, at a Southern Regional Graduate Summer Session held at the Virginia Polytechnic Institute in the summer of 1954: and a third time as a post-graduate course in the Michaelmas Term of 1954 and the Lent Term of 1955 at the London School of Economics. There has since been a fairly consistent demand for copies of the lecture notes, the original supply of which is now exhausted. It has therefore been decided to revise them and to issue them in the present form for the general use of students of statistics.

Multivariate Analysis in statistics is apt to be a baffling subject, especially for those students who want to use it in solving practical problems but do not possess the time or the inclination to plumb the depths of the mathematical theory to which it leads. This course was prepared with practical applications very much in the foreground. In it I have tried to expound the essential concepts and techniques and have limited the mathematical treatment as much as possible. In the present stage of knowledge this is no loss. The analysis of multivariate material requires to an unusual degree that peculiar blend of insight and skill in probabilistic interpretation which characterises the statistician, and for which pure mathematics is no substitute. It will, nevertheless, be evident that a considerable body of prerequisite knowledge is needed to get the most out of the course: mathematics up to matrix algebra, three-dimensional co-ordinate geometry and beta functions; statistical theory up to the theory of correlation and regression, the bivariate normal surface, and tests of significance based on normal theory.

I hope that the course may be found of some interest and use. Its issue in this form is experimental and I am indebted to the publishers, Messrs. Charles Griffin & Co., for the willingness they have shown to explore many possible methods of making available at a moderate price work which would be very costly to set in print.

M.G.K.

London,
August, 1957.

CONTENTS

INTRODUCTION 5

COMPONENT ANALYSIS : 10

Basis theorem on effective dimension-number 11

Principal components 13

Numerical solution of the characteristic equation 19

Acceleration by powering 24

The centroid method 27

A ranking approximation to the first component 35

FACTOR ANALYSIS : 37

The treatment of communalities 43

Estimating communalities 45

FUNCTIONAL RELATIONSHIP : 52

Berkson's case 55

Geary's case 59

Classical case 61

CANONICAL ANALYSIS 68

SOME PROBLEMS OF SAMPLING : 86

Significance and estimation in component and factor analysis 93

Estimation in factor-analysis models - Lawley's investigation 99

NOTES ON THE HISTORY OF MULTIVARIATE ANALYSIS : 105

Note on the history of Wilks' criterion 106

A historical note on latent roots 109

A historical note on discriminatory analysis 111

TESTS ON HOMOGENEITY : 117

The Pearson-Wilks results for bivariate populations 117

The analysis of dispersion 130

DISCRIMINATORY ANALYSIS : 144

The significance of a discriminant function 158

The case of k populations 163

REFERENCES 171

EXERCISES 180

A

COURSE IN

MULTIVARIATE ANALYSIS

1. INTRODUCTION

1.1 In a general sense "multivariate" analysis would include practically the whole of statistical theory. Even in so-called univariate problems, such as 'Student's' test of the mean, we require the idea of independence (of mean and variance in normal samples) and the complementary idea of dependence; and indeed any sample is a special case of a multiple variate in which the individuals, as a rule, are independent and identically distributed.

1.2 By general consent the term "multivariate analysis" is used in a much narrower sense. We find it easier to say what it is not that what it is, and many writers prescribe their domain of discussion by enumeration rather than by definition. When we look at the whole field, however, we discern two main features :

(a) We are concerned with a set of n individuals each of which bears the value of p different variates. The multivariate character, so to speak, lies in the multiplicity of the p variates, not in the size of the set n.

(b) The variates are dependent among themselves so that we cannot split off one or more from the others and consider it by itself. The variates must be

considered together.

1.3 Another important characteristic of multivariate analysis arises from a divergence of interest between the mathematician and the statistician. The natural inclination of the mathematician is towards generalization. Give him a result for one variate and he inquires after the result for two; give him that and he inquires after the result for p. The statistician, on the other hand, is continually struggling to reduce the dimensions of his problem. In multivariate analysis he usually has an embarrassing profusion of variates and his object is to make p as small as he can. (He still prefers n as large as possible.) Discriminant analysis, for example, tries to reduce the problem of distinguishing between multivariate populations to the scale of a single variate.

1.4 We may thus define multivariate analysis as the branch of statistical analysis which is concerned with the relationships of sets of dependent variates. We shall subdivide the main block of the subject into two parts, according to whether we are concerned with *dependence* or *interdependence*.

In *dependence*, one (or more) of the variates is selected for us by the conditions of the problem and we require to investigate the way in which it depends on the other variates – the so-called but badly-named "independent" variates. The regression of one variate on others or the variance analysis of a set of yields in a factorial experiment are of this type.

In *interdependence* we are concerned with the relationship of a set of variates among themselves, no one being selected as special in the sense of the dependent variate. The analysis of functional relationships, correlation and component analysis fall into this group.

1.5 Some parts of this field are covered at a comparatively early stage in statistical training: for example, partial association, partial correlation, the regression of one scalar variate on a set of others, and the analysis of variance. We shall not go over such ground again in this course. On the other hand, there are some portions which are rarely covered in quite advanced courses, such as component analysis and the

analysis of functional relationships. The reasons for this unbalanced scheme of instruction are partly historical and partly due to didactic convenience (and perhaps we ought to add, partly to the very severe theoretical problems which arise). The present course will attempt to exhibit the subject as a connected whole, though, naturally, more time will be spent on those branches which are less familiar to the student. Mathematical results will be proved or quoted, as seems convenient, but an attempt will be made to explain their significance so that the student who does not wish to delay over mathematical details can take the results on trust and return to their justification later. The emphasis throughout will be on practical applications. Computational methods will be taken for granted except (as in the case of component analysis) where they are of an unfamiliar type.

1.6 To give concreteness to the exposition we list here some types of practical problems with which multivariate analysis is concerned. Some, but not all, of these problems will be discussed in the sequel.

(a) *Biometrics* A number of skulls are dug up on an ancient burial ground. They may all come from one race or they may be a mixture of two opposing races, friend and foe having been flung into one pit together. An unlimited number of measurements can be made on any one skull. What are the best measurements to take, what is the minimum number we require and how do we use them to test homogeneity or heterogeneity in the sample ?

(b) *Education* A number of candidates n take an examination in p parts and are given a mark for each part. What is the best system of arranging the candidates in order of merit, and does the notion of "order of merit" have any justifiable meaning ?

(c) *Agriculture* A number of n different areas each produce yields of p different crops. Can we make any comparisons of general productivity between areas, and, if so, how ? Which crops are the best indicators of such a quality if it exists ?

(d) *Sociology* The replies given by members of a population to a questionnaire are expected to vary according to their social class. Information is collected about certain objective properties of a sample, e.g. rent, possession of a telephone, type of education. Can an index of social-class be constructed from such material ? And how should we test the significance of the difference between two samples ?

(e) *Medicine* A drug such as cortisone is injected into rats, which are subsequently killed and examined. Various organs are found to be affected as compared with a control group. How many of these effects are significant ? Which are the organs most affected ?

(f) *Physics* A set of plastics are tested for various physical properties such as resilience, strength, elasticity, ability to withstand abnormal temperatures. Can we detect in the results any systematic effects such as would enable us to predict them from known molecular properties of the plastics ? Can we then use the results to design better plastics for given purposes ?

(g) *Anthropology* For a set of Red Indian tribes there are recorded a number of items such as whether a tribe has a rain god, whether it uses totems, whether it has any agriculture. It is required to produce from this material criteria to decide whether a tribe belongs or not to certain ethnic groups, or to suggest what such groups might be.

(h) *Economics* Each year there is produced for a given country data which are in some way bound up with its general business activity, such as national income, rate of interest, freight-car loadings, steel production, unemployment, bank clearings and marriage rate. Can we produce an index-number of "business activity" from this complex, and if we can does it have any objective meaning ?

(i) *Experimentation* By accident or design a multi-factor experiment is conducted with no proper balance or orthogonal properties. How do we assess the significance of the main effects ?

(j) *Industry* A firm of tailors making ready-to-wear suits of clothes wishes to produce enough to cover the requirements of its large clientele with the minimum of misfits and unsold garments. The operative measurements are leg length, hip girth, trunk length, arm length, chest girth, shoulder width and perhaps a few others. On what measurements should it work and how should it proceed to produce maximum satisfaction at minimum cost ?

These examples, written down more or less at random, illustrate the wide field of application for multivariate analysis. This course is concerned with the methods which have been devised to investigate them.

2. COMPONENT ANALYSIS

2.1 Suppose we have p variates $x_1 \ldots x_p$, each observed on n individuals. We write x_{ij} for the jth observation on the ith variate so that the observations may be arrayed in a matrix:

$$\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ . & . & \ldots & . \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{bmatrix} \qquad (2.1)$$

The object of component analysis is to economise in the number of variates. To do this we shall seek for linear transformations of type

$$\zeta_i = \sum_{j=1}^{p} a_{ij} x_j, \qquad i = 1, 2, \ldots p. \qquad (2.2)$$

(We confine ourselves throughout to linear transformations. There is no reason why more complicated types should not be considered but the theory would become difficult. In practice, when the variation is obviously non-linear it is best to try to transform to linear variation before embarking on a component analysis.)

2.2 It may be that we can express the data in terms of fewer than p of the ζ's. We have then effected a genuine reduction in the dimensions of the problem, the whole complex of variations being expressible in $m < p$ variates. But this is exceptional. Where it is not possible we shall try to carry out an approximate reduction in this sense: we shall choose the coefficients a so that the first of our new variates ζ_1 has as large a variance as possible; we shall then choose the second

ζ_2 so as to be uncorrelated with the first and to have as large a variance as possible; and so on. In this way we transform to new uncorrelated variates (a useful thing in itself) which account for as much of the variation as possible in descending order. It may be that the first two or three of these variates account for "nearly" the whole of the variation, say 85 or 90 per cent, and the contribution of the other $p - 2$ or $p - 3$ is small. We can then say that the variation is represented *approximately* by the first two or three variates and in favourable circumstances may be able to neglect the remainder.

Basic theorem on effective dimension-number

2.3 We consider first of all the case when the number of ζ's is less than p; and we require the following results :

1. If the matrix (x_{ij}), $i = 1, \ldots p$, $j = 1, \ldots n$ is of rank m all the values x_{ij} are linearly dependent on m sets of them. In geometrical language, if we represent the n points in a Euclidean space of p dimensions by taking $x_1 \ldots x_p$ as co-ordinates, they will lie in a flat space of $m < p$ dimensions.

2. The rank of the product of a matrix by its transpose is equal to the rank of the matrix.

Now let us take each x_i measured about its mean, so that

$$\sum_{j=1}^{n} x_{ij} = 0, \quad i = 1, 2, \ldots p \tag{2.3}$$

and also let us standardize so that the x's have unit variance. Then

$$\frac{1}{n} \sum_{j=1}^{n} x_{ij}^2 = 1, \quad i = 1, 2, \ldots p. \tag{2.4}$$

Then the $p \times p$ matrix whose i, jth term is

$$\operatorname{cov}(x_i, x_j) = \frac{1}{n} \sum_{k=1}^{n} x_{ik} x_{jk}$$

is a factor in n times the product of (x_{ij}) and its transpose. Except for a factor in n this is the correlation matrix

$$(r_{jk}) = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{bmatrix} \quad (2.5)$$

Then it follows that: if and only if the correlation matrix is of rank $m < p$ the variation lies in a linear space of m dimensions (and consequently it is possible to find a linear transformation to m new variates which completely account for the variation).

Example 2.1

Consider a four-variate case with matrix

$$\begin{bmatrix} 1 & .8 & .6 & .6 \\ .8 & 1 & .96 & 0 \\ .6 & .96 & 1 & -.28 \\ .6 & 0 & -.28 & 1 \end{bmatrix}$$

An examination of the matrix (e.g. the calculation of its determinants and of the determinants of its three-rowed minors shows that it is of rank 2 ($p = 4$, $m = 2$). It is therefore possible to find a linear transformation to two new variates ζ_1 and ζ_2 and only two variates are needed to express the data. One set would be given by

$$\begin{aligned} x_1 &= \zeta_1 \\ x_2 &= 0.8\,\zeta_1 + 0.6\,\zeta_2 \\ x_3 &= 0.6\,\zeta_1 + 0.8\,\zeta_2 \\ x_4 &= 0.6\,\zeta_1 - 0.8\,\zeta_2 \end{aligned}$$

but there are others.

Example 2.2

Consider the matrix ($p \times p$)

$$\begin{bmatrix} 1 & r & r & \ldots & r \\ r & 1 & r & \ldots & r \\ . & . & . & \ldots & . \\ r & r & r & \ldots & 1 \end{bmatrix}$$

Adding the rows and taking out a factor in $\{1 + (p - 1)r\}$, and then subtracting r times the unit row from all other rows, we see that the determinant of this matrix is

$$(1 - r)^{p-1} \{1 + (p - 1)r\} .$$

This cannot vanish unless $r = 1$ or $r = 1/(p - 1)$. Except in these special cases the rank of the matrix must then be p and hence we cannot represent a set of equally correlated variates in fewer than p dimensions.

2.4 It is of some interest to record a result concerning the rank of a *symmetric* matrix which relieves us of the necessity for testing every minor: if one m-rowed principal minor is not zero and if that minor vanishes (a) when any one row and the corresponding column is annexed to it and (b) when any two rows and the corresponding columns are annexed to it, the rank is m.

There are p rows and a row can be annexed in $p - m$ ways and two rows in $\frac{1}{2}(p - m)(p - m - 1)$ ways. The number of conditions on a symmetric matrix for it to be of rank m is then $\frac{1}{2}(p - m)(p - m + 1)$. These are, in fact, independent conditions. (Cf. Lederman, W., 1937, *Psychometrika*, 2, 85).

Principal components

2.5 Consider the n points in the space of p dimensions when the x's are expressed in standard measure (zero mean, unit variance). Consider the line with current co-ordinates X

$$\frac{X_1 - m_1}{l_1} = \frac{X_2 - m_2}{l_2} = \ldots = \frac{X_p - m_p}{l_p} \qquad (2.6)$$

where the l's are direction cosines and are therefore subject to the condition

$$\sum_{i=1}^{p} l_i^2 = 1. \qquad (2.7)$$

The sum of squares of the distances from the n points on to this line is nS, say, given by

$$nS = \sum_{j=1}^{n} \left[\sum_{i=1}^{p} (x_{ij} - m_i)^2 - \{ \sum_{i=1}^{p} l_i (x_{ij} - m_i) \}^2 \right] \qquad (2.8)$$

If this is a stationary value the partial differentials with respect to m's vanish and hence

$$- \sum_j (x_{ij} - m_i) + \sum_j l_i \sum_i l_i (x_{ij} - m_i) = 0, \quad i = 1, 2, \ldots p. \qquad (2.9)$$

and since $\sum_j x_{ij} = 0$ this leads to

$$\frac{m_i}{l_i} = \text{constant}.$$

Hence the origin lies on the line (2.6) and without loss of generality we may take all the m's to be zero. Then, since $\sum_{j=1}^{n} x_{ij}^2 = n$, we have

$$nS = \sum_{j=1}^{n} \left[\sum_{i=1}^{p} x_{ij}^2 - (\sum_{i=1}^{p} l_i x_{ij})^2 \right]$$

$$= np - \sum_{j=1}^{n} (\sum_{i=1}^{p} l_i x_{ij})^2. \qquad (2.10)$$

We then find the stationary values of S for variations in l subject to (2.7). If λ is an undetermined multiplier this leads to

$$- \frac{1}{n} \sum_{j=1}^{n} x_{kj} (\sum_i l_i x_{kj}) + \lambda l_k = 0, \quad k = 1, 2, \ldots p \qquad (2.11)$$

or the set of p equations

$$\begin{aligned} l_1(1-\lambda) &+ l_2 r_{12} &&+ \ldots + l_p r_{1p} = 0 \\ l_1 r_{21} &+ l_2(1-\lambda) &&+ \ldots + l_p r_{2p} = 0 \\ \cdot &\quad \cdot &&\quad \cdot \\ l_1 r_{p1} &+ l_2 r_{p2} &&+ \ldots + l_p(1-\lambda) = 0 \end{aligned} \qquad (2.12)$$

If we eliminate the l's we get the so-called characteristic equation of the correlation matrix which may be written

$$| r - \lambda I | = 0 . \qquad (2.13)$$

For known r this gives us, in general, p roots in λ. To each root corresponds a set of l's for which S has a stationary value. Moreover, by using (2.11) we find from (2.10)

$$S = p - \lambda . \qquad (2.14)$$

λ cannot be negative.

Furthermore, it follows from (2.14) that the root which gives the minimum S is the one with the largest λ. Choosing the largest root of (2.13) therefore gives us the line we require. The sum of squares of distances of the points from it is a minimum and the variate measured along it has the maximum variance. The variate is given by

$$\zeta_1 = \sum_{j=1}^{p} l_{1j} x_j \qquad (2.15)$$

where we write l_{1j} with the subscript 1 to indicate that this set of l's relate to λ_1. Multiplying (2.11) by x_k and summing over k we see that the variance of ζ_1 is λ_1.

2.6 It is not immediately obvious, that if we now seek for the direction (perpendicular to our line) for which the sum of squares of perpendiculars is a minimum we shall arrive at the line corresponding to λ_2, the second largest root of (2.13); and so on. But such is the fact. Let us prove in the first place that if $l_{\alpha i}$, $l_{\beta i}$ are the l's corresponding to any two

different roots λ_α and λ_β then the l's are orthogonal, that is to say

$$\sum_{i=1}^{p} l_{\alpha i}\, l_{\beta i} = 0, \qquad \alpha \neq \beta . \tag{2.16}$$

In fact, from (2.12) we have

$$\sum_j l_{\alpha j}\, r_{ij} = \lambda_\alpha\, l_{\alpha i},$$

$$\sum_j l_{\beta j}\, r_{ij} = \lambda_\beta\, l_{\beta i}.$$

Multiply the first by $l_{\beta i}$ and the second by $l_{\alpha i}$, sum over i and subtract. We have

$$\sum_{i,j} (l_{\beta i}\, l_{\alpha j}\, r_{ij} - l_{\alpha i}\, l_{\beta j}\, r_{ij}) = \sum_i (\lambda_\alpha\, l_{\alpha i}\, l_{\beta i} - \lambda_\beta\, l_{\beta i}\, l_{\alpha i}).$$

Since $r_{ij} = r_{ji}$ the left-hand side vanishes and thus

$$(\lambda_\alpha - \lambda_\beta) \sum l_{\alpha i}\, l_{\beta i} = 0,$$

from which (2.16) follows unless $\lambda_\alpha = \lambda_\beta$.

Thus if $\zeta_1, \ldots \zeta_p$ correspond to the roots $\lambda_1, \ldots \lambda_p$ the lines form an orthogonal set. In geometrical terms we have rotated our axes of co-ordinates from the x - set to the ζ set. The matrix (l_{ij}) is self-orthogonal. It follows that since

$$\zeta_i = \sum_j l_{ij}\, x_j$$

we have also

$$x_i = \sum_j l_{ji}\, \zeta_j. \tag{2.17}$$

Also

$$\operatorname{cov}(\zeta_i, \zeta_j) = \operatorname{cov}\Big(\sum_k l_{ik}\, x_k \sum_m l_{jm}\, x_m\Big)$$

$$= \sum_{k,m} l_{ik}\, l_{jm}\, r_{km}$$

$$= \lambda_i \sum_m l_{im}\, l_{jm}$$

$$= 0 \quad \text{unless} \quad i = j,$$

and hence the variates ζ are statistically uncorrelated.

2.7 We have thus transformed to new variates ζ which are uncorrelated and have variances $\lambda_1, \lambda_2, \ldots \lambda_p$ in decreasing order. We note that $p = \Sigma\lambda$ as is also evident from (2.13); for the sum of the roots of a p-ic in λ is the sum of p units in the main diagonal.

In particular, if the variates are normally distributed we may regard the ζ's as splitting off independent components of variance $\lambda_1, \lambda_2, \ldots \lambda_p$ from the total p.

Note the following points which we state without proof:

(a) All roots of the characteristic equation (2.10) are real and non-negative. This is a property of non-negative definite matrices such as the correlation matrix.

(b) In degenerate cases certain λ's may be equal; they may even be all equal. This is not of much practical importance. If it occurs the problem of determining a unique line or lines minimizing the sum of squares is indeterminate and an infinite set will satisfy the conditions.

(c) If certain λ's vanish, say $p - m$ of them, the correlation matrix is of rank m and we are back to the case of 2.3 in which the variation collapses into a space of m dimensions. The last $p - m$ ζ's are then not required.

(d) We have standardized the variates by reducing them to unit variance before finding the ζ's. Had we not

done so, but left the scales unchanged, we should have found different ζ's which are not transformable into our set by standardization after the new variates are determined. In geometrical language, lines of closest fit found by minimizing sums of squares of perpendiculars are not invariant under change of scale. This point is troublesome when we consider sampling problems.

Example 2.3

(Kendall, 1939, *J. Roy. Statistical Soc.*, 102, 21.)

The yields of ten crops were recorded for 48 counties in England. ($n = 48$, $p = 10$). The crops were wheat, barley, oats, beans, peas, potatoes, turnips, mangolds, hay (temporary grass) and hay (permanent grass). The correlations between the various crops were nearly all positive, suggesting that there might be some quality "productivity" associated with an area irrespective of the crops actually grown. To allow for climatic variations four years were chosen; the results for them agreed quite closely.

The correlation matrix was computed and the largest root ascertained. For 1925, for instance, $\lambda = 4.760$ and thus the corresponding ζ accounts for 47.6 per cent of the total variation. The corresponding ζ was given by

$$\begin{aligned}\zeta = {} & 0.39x_1 + 0.37x_2 + 0.39x_3 + 0.27x_4 + \\ & 0.22x_5 + 0.30x_6 + 0.32x_7 + 0.26x_8 + \\ & 0.24x_9 + 0.34x_{10}. \end{aligned} \qquad (2.18)$$

This variate was provisionally identified with productivity and the counties were arranged in order according to the magnitude of ζ as given by (2.18). They were then grouped into very good, good, moderate, poor, bad according to the values of ζ which they bore. The results agreed with general knowledge about the geographical distribution of productivity except in one or two instances.

In this case a value of $\lambda = 4.76$ is not very high and we

suspect that another variate, at least, is required to "explain" the variation. A more detailed analysis is given by Banks, C.H. (1954), *J. Roy. Statist. Soc.* B. We note that all coefficients in (2.18) are positive.

From one point of view (2.18) may be regarded as determining an index ξ of productivity. In the ordinary way one might expect the crop yields to be weighted in some way according to the production or acreage of the various crops. In the 1939 paper Kendall tried alternative methods weighting both by value and by starch-equivalent but the results were much the same as by the use of (2.18).

The problem here was to reduce the 10-dimensional variation to a one-dimensional variation so that the resulting variate might be used as a measure of productivity. The attempt was partly successful, and as successful as any attempt can be which endeavours to express productivity as due to the variation of one component only.

Numerical solution of the characteristic equation

2.8 The best method of solving the characteristic equation is by iteration. It has the advantage that the l's corresponding to any root are found simultaneously. We will illustrate on the matrix of example 2.1, which we already know to contain two components only:

				A	B
1	.8	.6	.6	1.0	1.0000
.8	1	.96	0	0.92	.9974
.6	.96	1	−0.28	0.76	.8632
.6	0	−0.28	1	0.44	.3368
3.0	2.76	2.28	1.32		(2.19)

We add the columns as shown, divide by the largest and write vertically as at column A. We multiply the rows of the matrix by the figure in the corresponding row in A to obtain a new matrix

1.0000	.8000	.6000	.6000	
.7360	.9200	.8832	.0000	
.4560	.7296	.7600	− .2128	
.2640	.0000	−.1232	.4400	
2.4560	2.4496	2.1200	0.8272	(2.20)

Again we sum columns, divide by the sum in the same column as before and write the results at B in (2.19). We repeat the process until we get two columns of types A and B identical. At that stage we have the l's corresponding to the largest λ given as proportional to the numbers in the final column. Thus, after 12 iterations on (2.19) we find 1.0000, 1.10000, 1.0000, .2000. (This, as it happens, is exact.) These are proportional to the l's. The sum of their squares is 3.25 and hence the l's are

$$\frac{1}{\sqrt{3.25}} \quad (1,\ 1.1,\ 1,\ 0.2) \qquad (2.21)$$

The corresponding λ is given by multiplying these figures by the corresponding terms in the first row of the matrix and dividing by l_1, giving

$$\lambda = \frac{1}{\sqrt{3.25}} \{(1 \times 1) + (1.1 \times .8) + (1 \times .6) + (.2 \times .6)\} \Big/ \frac{1}{\sqrt{3.25}}$$

$$= 2.6 .$$

2.9 The validity of the process is evident from equation (2.9). Since the correlation matrix is symmetrical the first stage of addition by columns is equivalent to trying $l_1 = l_2 = \ldots = l_4$ proportional to unity in the equations of type

$$\Sigma l_i r_{ij} = l_j \lambda$$

and the sums are, as a second approximation, proportional to the right-hand side, namely to the l's. Using these new values we obtain a third approximation, and so on until the

process converges. We now prove (a) that it always will converge if the λ's are different and (b) that unless by accident we happen to hit on a case where the first trial is exactly a root corresponding to a smaller λ (a possibility so remote as to be negligible) the process will converge to the largest λ.

In fact, consider a line whose direction cosines in the original space are $a_1, a_2 \ldots a_p$. If we transform to the ξ-axis the direction cosines become, say, $b_1, b_2, \ldots b_p$ related to the a's by

$$a_j = \sum_k l_{jk} b_k.$$

If the a's relate to an approximation to the direction cosines of a principal axis and a to the next approximation determined by

$$\Sigma r_{ij} a_j = a_i'$$

we shall then have the corresponding relation

$$\sum_{j,k} r_{ij} l_{jk} b_k = a_i' = \sum_m l_{im} b_m.$$

But

$$\sum_j r_{ij} l_{jk} = \lambda_k l_{ik}$$

and hence

$$\Sigma \lambda_k l_{ik} b_k = \Sigma l_{im} b_m'$$

or

$$\Sigma (\lambda_k b_k - b_k') l_{ik} = 0.$$

This is equivalent to

$$b_k' = \lambda_k b_k.$$

Now if we start with trial a's and construct successive approximations in the manner described, this is equivalent, in the co-ordinate system of the ξ's, to the successive application

of $b'_k = \lambda_k b_k$. If, moreover, $\lambda_1 > \lambda_2 > \ldots \lambda_p$, each of the ratios b'_j / b'_1 is less than b_j / b_1 unless $b_1 = 0$. Under iteration these ratios therefore approach zero. Hence

(a) In general the direction cosines with respect to the r-system tend to zero except for b_1 and hence the process converges to the largest root and the corresponding principal axis;

(b) This fails if $b_1 = 0$, in which case the successive approximations all give lines perpendicular to the major principal axis and the process converges to the largest root in the hyperplane perpendicular to that axis;

(c) If several of the b's, e.g. b_1, b_2, b_3 are equal (corresponding to equal λ's) the remaining direction cosines tend to zero with respect to them; but for the equal l's the process does not converge further in the sense of differentiating between them.

2.10 Having found the largest root we proceed to find the next largest. To do so we "extract" from the original matrix the first component in the following manner :

Form the matrix $\lambda\, l_i\, l_j$ to get

.8	.88	.8	.16
.88	.968	.88	.176
.8	.88	.8	.16
.16	.176	.16	.032

(2.22)

Subtract this from the original matrix to get

.2	−.08	−.2	.44
−.08	.032	.08	−.176
−.2	.08	.2	−.44
.44	−.176	−.44	.968

(2.23)

Then proceed to apply the same procedure as before to this residual matrix. (We justify this procedure below).

In this instance we find for the sums of columns in (2.23): 0.36, −0.144, − 0.36, 0.792 and the iteration converges at once to l's proportional to 1, −0.4, −1, 2.2 with $\lambda = 1.4$. If we go on to extract this from (2.23) we find a vanishing matrix and hence we have exhausted the variation; this is confirmed by the fact that our two λ's, 2.6 and 1.4 add up to 4, the value of p. The second and final component has then l's equal to

$$\frac{1}{\sqrt{7}}(1, -0.4, -1, 2.2) \qquad (2.24)$$

with $\lambda = 1.4$.

2.11 The procedure of "extracting" components from a matrix in the manner described may be justified as follows.

The transformation to new variable ζ results in a set of variables which are uncorrelated. We have

$$\begin{aligned} \text{cov}\,(x_i, x_j) &= E\{\Sigma\, l_{ki}\,\zeta_k \;\Sigma\, l_{mj}\,\zeta_m\} \\ &= \Sigma\, l_{ki}\, l_{mj}\, E\,(\zeta_k \zeta_m) \\ &= \Sigma\, l_{ki}\, l_{kj}\, \lambda_k. \end{aligned} \qquad (2.25)$$

If we now fix ζ_1 the covariance of the resulting conditioned x_i and x_j is the sum on the right in (2.25) less the first term $l_{1i}\, l_{1j}\lambda_1$. Thus the covariance after the first component is removed is given by the procedure of 2.10, for the original matrix (by reason of standardization) is a covariance matrix.

To find the next component we operate on this residual matrix but we do not re-standardize it so that the diagonal terms are unity. Geometrically viewed, the determination of the second component is equivalent to finding the line, in the space perpendicular to the first component, for which the sums of squares of distances is a minimum. It is not, perhaps, obvious that this will give us the second largest root of the

characteristic equation (2.13); but, as we noted in 2.6, the λ_2-direction is perpendicular to that corresponding to λ_1 and if the sum of squares is stationary the corresponding direction must coincide with the λ_2-direction. The original covariances are reduced by terms like $l_{ki}\, l_{kj}\, \lambda_k$ at the extraction of the kth factor, as we see from (2.25).

Acceleration by powering

2.12 The convergence procedure of 2.8 may be greatly accelerated by the following device: reverting again to the original matrix (2.19), square it to get

2.3600	2.1760	1.8000	1.0320
2.1760	2.5616	2.4000	0.2112
1.8000	2.4000	2.3600	−0.2000
1.0320	0.2112	−0.2000	1.4384
7.3680	7.3488	6.3600	2.4816

(2.26)

If we now form the column sums and proceed as before our first trial column is 1.0, .9974, .8632, .3368, which is the second column of (2.19). We should find that the second trial column of (2.26) would be the fourth of (2.19); and, generally, that the convergence is twice as fast.

Nor need we stop here. If we square (2.26) and operate on the resultant

14.6096	15.2474	13.5120	4.0195
15.2474	17.1014	15.6864	2.6104
13.5120	15.6864	14.6096	1.6048
4.0195	2.6104	1.6048	3.2186
47.3885	50.6456	45.4128	11.4533

(2.27)

and the first trial column gives

1.0	1.0687	0.9583	0.2417

which is the same as we should get by four iterations on the original matrix. Generally, if we raise the matrix to the tth power the convergence is t times as fast. (Convenient numbers for practice are $t = 4$ or $t = 8$. It does not pay to power too far because the elements of the matrix then contain too many digits.)

2.13 The reason for this is easily seen. If l'_j is a trial set of l's so that the next set is

$$l''_i = A \sum_j l'_j r_{ij}$$

where A is a constant, the next set is

$$l'''_k = B \sum_j l''_j r_{ij} = B \sum_j r_{jk} \sum_i l'_i r_{ji}$$

$$= B \sum_i l'_i \sum_j r_{kj} r_{ji}$$

and the second summation on the right is r_{ij} squared.

2.14 Other methods of extracting roots are known. Burt (1937, *Brit. J. Ed. Psych.* **7**, 172) points out that in the powered matrix the diagonal elements tend to become proportional to the squares of the l's. Aitken (1937, *Proc. Roy. Soc. Ed.* A, **37**, 269) proceeds by pivotal condensation. Anyone who has much of this kind of arithmetic to do is probably well advised to consult an expert. Most electronic machines now carry a programme which will extract the characteristic roots λ and the corresponding vectors l_{ij}.

2.15 Note one possible source of confusion. As we have defined the components ζ they do not have unit variances. In our example the new variates, from (2.21) and (2.23) are

$$\zeta_1 = \frac{1}{\sqrt{3.25}} \{x_1 + 1.1\, x_2 + x_3 + 0.2\, x_4\} \qquad (2.28)$$

$$\zeta_2 = \frac{1}{\sqrt{7}} \{x_1 + 0.4\, x_2 + x_3 + 2.2\, x_4\} \qquad (2.29)$$

with variances 2.6 and 1.4 respectively. (We note immediately that they are orthogonal, as they must be.) In factor analysis it is more usual to standardize the ζ's so that they themselves have unit variance. Thus we should have two "factors", say f_1 and f_2, such that $\zeta_1 = \sqrt{(2.6)} f_1$ and $\zeta_2 = \sqrt{(1.4)} f_2$. On substitution in (2.28) and (2.29) this gives us

$$f_1 = \frac{1}{\sqrt{8.45}} \{x_1 + 1.1\, x_2 + x_3 + 0.2\, x_4\}$$

$$= .3440\, x_1 + .3784\, x_2 + .3440\, x_3 + .0688\, x_4 \qquad (2.30)$$

and

$$f_2 = \frac{1}{\sqrt{9.8}} \{x_1 - 0.4\, x_2 - x_3 + 2.2\, x_4\}$$

$$= .3194\, x_1 - .1278\, x_2 - .3194\, x_3 + .7028\, x_4 \qquad (2.31)$$

Example 2.4

(Stone, 1947, *Supp. J. Roy. Statist.* Soc. 9, 1)

Stone took Kuznet's and Barger's data for the U.S.A. comprising, for each of the years 1922–1938, 17 series regarded as constituent elements of total national income or outlay, e.g. employers' compensation, consumers' perishable goods plus producers' durable goods, net public outlay, net increase in inventories, dividends, interest, foreign balance and so on. He did a principal component analysis on the observations taken about their mean and extracted three principal components with λ = 80.76, 10.59, 6.09 per cent, accounting for 97.45 per cent of the variance. Evidently these three components account for nearly all the variation and we have thus reduced the effective dimensions of variation from 17 to 3. This illustrates the economy in effective dimension number which is the object of component analysis to achieve.

The remarkable feature of Stone's work, however, is that he was able to interpret his components. In many cases our principal components do not have an identifiable separate

existence and are to be regarded as convenient mathematical artefacts. In others (the "general intelligence" factor is a notorious case) it is arguable whether the components can be given any reality. Stone, however, had reason to suppose on economic grounds that the variation was mostly accounted for by three components (a) total income i or some similar quantity, (b) rate of change of i, say Δi and (c) a trend term expressing expansion or contraction of the economy, which we may take as a linear term in the time t. Moreover, these quantities were separately measurable. Stone correlated his three principal components, say F_1, F_2, F_3 (in his notation) with i, Δi and t to obtain

	F_1	F_2	F_3	i	Δi	t
F_1	1					
F_2	0	1				
F_3	0	0	1			
i	.995	−.041	.057	1		
Δi	−.056	.948	−.124	−.102	1	
t	−.369	−.282	−.836	−.414	−.112	1

The three underlined figures stand out and it seems very reasonable to identify F_1 with i, F_2 with Δi and F_3 with t.

The results of the inquiry are to be interpreted cautiously because there were only 17 observations, but the study is a remarkable one. Stone notes that in demand analysis, when including a price we ought to include a large number of other prices. This is usually impossible, but we may be able to introduce a principal component or two representing the price-complex through an index number.

The centroid method

2.16 Psychological workers have developed numerous methods of component analysis which avoid the arithmetic required by the

solution of the characteristic equation. My personal opinion is that they are objectionable and should not be used when they can be avoided. So much published work has been based on them, however, that some account is necessary. Perhaps they can be justified to some extent as giving approximations to the principal component method, but any discussion of their sampling properties seems almost beyond the range of reasonable possibility.

2.17 The matrix of observations (2.1) can be regarded as defining not only n points in p dimensions but p points in n dimensions. Actually these p points, one corresponding to each variate, will, together with the origin, define a p-dimensional space embedded in the n-space. We may imagine them as radiating from the origin like the spokes of an umbrella. Their lengths are proportional to the variances (in our case all unity) and the cosines of the angles between them are the correlation coefficients.

2.18 If the correlations r_{ij} are large the angles between the vectors are small and they bunch together. If there is a common component we may expect it to go somewhere through the "middle" of this bunch. The centroid method formalises this idea by supposing that the first component passes through the centre of gravity of the points.

Let us take a set of orthogonal co-ordinate axes in this second kind of p-dimensional space. If our original x's have been standardized so as to have unit variance the p points in this space will be at the ends of p unit vectors radiating from the origin. Let the co-ordinates of the ith point in the jth dimension be y_{ij}. Then

$$\sum_{j=1}^{p} y_{ij}^2 = 1$$

and

$$r_{ij} = \sum_{k=1}^{p} y_{ik} y_{jk}. \qquad (2.32)$$

The centroid of the p points is then given by $y_{.j}$, say, where

$$y_{.j} = \frac{1}{p} \sum_{i=1}^{p} y_{ij} \qquad (2.33)$$

Suppose we take the axis of the first co-ordinate to go through the centroid, as we may without loss of generality. Then $y_{.k} = 0$ unless $k = 1$. Hence

$$r_{.j} = \frac{1}{p} \sum_{i=1}^{p} r_{ij} = \Sigma y_{.k} y_{jk} = y_{.1} y_{j1} \qquad (2.34)$$

Let us sum the columns in the correlation matrix, giving us sum $A_1 \ldots A_p$ such that

$$A_j = pr_{.j} = p\, y_{.1} y_{j1}$$

and
$$T = \sum_j A_j = p^2 y_{.1} y_{.1} = (py_{.1})^2 \qquad (2.35)$$

This gives us
$$y_{.1} = \frac{1}{p} \sqrt{T}$$

and
$$y_{j1} = \frac{pr_{.j}}{\sqrt{T}} = \frac{A_j}{\sqrt{T}} \qquad (2.36)$$

Thus the first component, going through the centroid, has elements proportional to $A_j / \sqrt{T}$, which are easily found from the correlation matrix.

Example 2.5

Consider again the matrix of Example 2.1 :

1	.8	.6	.6
.8	1	.96	0
.6	.96	1	−.28
.6	0	−.28	1
Sum A_j	3.0	2.76	2.28	1.32

We find $$T = \Sigma A_j = 9.36,$$

$$1/\sqrt{T} = .326,\ 860,$$

and hence the first component is proportional to

$$\frac{1}{\sqrt{T}} \Sigma A_j x_j = .9806\, x_2 + .9021\, x_2 + .7[illegible]\, x_3 + .4315\, x_4 \ . \quad (2.37)$$

To make this comparable with the first principal component ξ_1 we need to standardize it so that the sum of squares of coefficients is unity. The sum of squares in (2.37) is 2.516, 876. On division by the square root of this we get for the first centroid component

$$.6181\, x_1 + .5686\, x_2 + .4697\, x_3 + .2720\, x_4 \ . \quad (2.38)$$

as against the first principal component (2.21)

$$.5548\, x_1 + .6163\, x_2 + .5548\, x_3 + .1110\, x_4 \ . \quad (2.39)$$

The variance of the linear function (2.37) is easily seen to be $\Sigma A_j A_k r_{jk} / T$ which in this case turns out to be 6.427,692. Division of this by 2.516,876 gives us, for the variance of (2.38), 2.554. This is to be compared with the optimum value 2.6 for the first principal component.

It should be noted that the first centroid component is equivalent to the first approximation of (2.19).

2.19 The determination of the first component is comparatively simple, but the real trouble lies ahead. If, in the previous example, we form the matrix of (2.37) :

.9616	.8846	.7307	.4231
.8846	.8138	.6722	.3893
.7307	.6722	.5553	.3216
.4231	.3893	.3216	.1862

(2.40)

and subtract from the original matrix, we get a residual

.0384	−.0846	−.1307	.1769	
−.0846	.1862	.2878	−.3893	
−.1307	.2878	.4447	−.6016	
.1769	−.3893	−.6016	.8138	
.0000	.0001	.0002	−.0002	(2.41)

We will prove presently that it is the matrix of (2.37) which is to be extracted, not, for example, that of (2.33). Taking this temporarily for granted, let us note that we can now proceed to extract a second component by the same method because all the column sums are zero (within errors of rounding up). And this must always be so. Geometrically, in the p-space we have found a vector passing through the centroid, and in (2.41) we have, in effect, projected the complex of vectors on to a space perpendicular to this vector. The centroid of the projected points obviously lies at the origin and we cannot draw a new vector through this origin and a new centroid.

2.20 We therefore proceed by reflecting one or more of the vectors in the origin so as to "break the centroid away" from the origin. Unfortunately, this involves an element of subjective judgement. Several rules have been proposed, e.g. to reflect the vectors with the most negative signs in the residual matrix. But there is no absolute rule.

Example 2.6

In example 2.5 the procedure is fairly clear. The signs in (2.41) run

+	−	−	+
−	+	+	−
−	+	+	−
+	−	−	+

We then "reflect" the vectors corresponding to x_2 and x_3. This will change the signs of the terms in the second and third columns and *again* those in the second and third rows, leaving the terms in both unchanged. The terms in the transformed matrix are then all positive and the sum of terms in the columns of (2.39) is

.4306 .9479 1.4648 1.9816

We find $T = 4.8249$, $1/\sqrt{T} = .455,256$. The second component has then coefficients proportional to

.1960 .4315 .6669 .9021.

But we must now "put the signs back" by changing those of the second and third terms. The second component is thus proportional to

$$.1960\, x_1 - .4315\, x_2 - .6669\, x_3 + .9021\, x_4 \qquad (2.42)$$

It is readily verified that this is independent of the factor (2.37).

2.21 The following points may be briefly noted :

(a) the sum of cross products in (2.37) and (2.40), namely

(.9806 x .1960) + (.9021 x −.4315) + (.7452 x −.6669)

+ (.4315 x .9021)

does not vanish. The centroid components are not orthogonal in the p-space which we considered in arriving at the principal components. They are orthogonal in our second p-space. Only the principal components are orthogonal in both.

(b) If we have to proceed to determine third, fourth components, etc., we do not put the signs back until the very end of the operation.

(c) If we are to proceed as far as p components any

method of "breaking the centroid" from the origin will do in the sense that we get p independent components. But they will not in general have any optimal properties in regard to variance.

2.22 It remains to justify the process of extracting the successive components in centroid analysis. We require a result similar to that of 2.11.

Suppose that, in some space of dimension p, we have a set of m vectors y_{ij} ($i = 1, 2, \ldots m$; $j = 1, 2, \ldots p$) (not necessarily of unit length) and that their covariances are typified by e_{ij}. If we project them perpendicularly to a unit vector t_i and

$$a_i = \sum_{j=1}^{p} y_{ij} t_j \qquad i = 1, \ldots m \qquad (2.43)$$

then the covariances of the projected vectors are given by e'_{jk} where

$$e'_{jk} = e_{jk} - a_j a_k \qquad (2.44)$$

Let CX, CY be two vectors of lengths d_j, d_k projected perpendicularly to a line CD on to a plane through D and meeting it in A, B respectively. Then

$$AB^2 = AD^2 + DB^2 - 2AD.\ DB \cos ADB$$

$$= AC^2 + BC^2 - 2AC.\ BC \cos ACB$$

In conjunction with the relations AD = AC sin ACD, BD = BC sin BCD this gives us

$$\sin ACD \sin BCD \cos ADB = \cos ACB - \cos ACD \cos BCD. \qquad (2.45)$$

But we also have by definition

$$e_{jk} = d_j d_k \cos ACB$$

and

$$e'_{jk} = AD.DB \cos ADB$$

From (2.45) we then derive

$$e'_{jk} = e_{jk} - d_j \cos \text{ACD}\, d_k \cos \text{BCC}$$

and since $a_j = d_j \cos \text{ACD}$, $a_k = d_k \cos \text{BCD}$ this results in (2.44).

To justify the process of successive extraction we need to note (1) that the reflection of vectors does not affect the dimensions of the space in which they lie and (2) that if there are m vectors y lying in a space of dimension p with covariances e_{jk} and we form

$$a_j = \frac{\sum_k e_{jk}}{\sqrt{(\sum_{j,k} e_{jk})}} \tag{2.46}$$

then there is a unit vector t_i lying in the p-space such that cov $(y_i, t) = q_i$.

For if we form the dispersion matrix of the y's and t

$$\begin{bmatrix} e_{jk} & a_j \\ a'_k & \sqrt{\Sigma a_{jk}} \end{bmatrix} \tag{2.47}$$

from the definition of the a's in (2.46) we see that this matrix has the same rank as e_{jk}. This proves that the first centroid component lies in the p-space of variation. When we extract it we project the vectors on to a plane orthogonal to this component. Any subsequent components are therefore uncorrelated with this component. The residual covariance matrix by (2.44) is the residual as we have calculated it. Reflection moves the centroid away from the origin and so we can prceed to the analysis of subsequent components.

A ranking approximation to the first component

2.23 If our object is to arrange the individuals in some sort of order according to a principal component (as, for example, if we wish to arrange students in order of a putative general intelligence as shown by test performances) a very fair approximation can often be obtained by simple ranking methods. For any variate we rank the n individuals from 1 to n. We then add the p ranks for each individual and so arrive at a score for him. The ordering of this score will give us a ranking of the individuals according to a common component.

Such a procedure, in fact, maximizes the average Spearman correlation between the ranking so reached and the rankings according to the p variates (Kendall, 1955, *Rank Correlation Methods*, 7.10). The rank-vector therefore gets as close to the p rank-vectors as it can, so to speak, and will approximate to the first centroid component and to the first principal component.

Example 2.7

In the case referred to in Example 2.3 Kendall ranked the 48 counties according to the yields for each of the 10 crops, summed the ranks and compared the order with that given by the first principal component. The agreement was strikingly close and a Spearman ρ between the orders given by the two methods was 0.99.

Stamp (1952, Land for Tomorrow, Bloomington Press, Indiana) has recently applied this idea to an agricultural grading of certain countries. The average ranks on nine crops (wheat, rye, barley, oats, corn, potatoes, sugar beet, beans and peas) were as follows :

	1934-8		1946
Belgium	2.2	Belgium	2.3
Denmark	2.6	Denmark	2.4
Netherlands	2.9	Netherlands	2.4
Germany	4.3	New Zealand	4.2
Britain	4.7	Britain	4.8

	1934–8		1946
Ireland	4.7	Ireland	5.3
New Zealand	5.8	Egypt	6.2
Egypt	6.3	Germany	7.6
Austria	7.2	U.S.A.	8.2
France	9.2	France	9.0
Japan	10.4	Canada	9.1
Italy	12.0	Austria	11.2
U.S.A.	12.0	Chile	11.5
Canada	12.3	Argentina	12.4
Spain	12.6	China	12.7
Chile	12.9	Italy	12.7
China	13.6	Japan	14.1
Argentina	14.3	Spain	14.2
Australia	16.0	India	17.0
India	17.8	Australia	17.2

With all its imperfections, this is an interesting table. The lowering of the positions occupied by countries which are supposed to have lost the war is noticeable; and the rise in countries remote from the actual battlefields, New Zealand, U.S.A., Canada and Argentina (but not Australia or India) is also brought out.

3. FACTOR ANALYSIS

3.1 In component analysis we begin with the observations and look for components in the hope that we may be able to reduce the dimensions of variation and also that our components may, in some cases, be given a physical meaning. We work from the data toward a hypothetical model. In factor analysis we work the other way round; that is to say, we begin with a model and require to see whether it agrees with the data and, if so, to estimate its parameters.

3.2 This is a broad distinction but it is often blurred in practice because, at different stages of the development of a subject, we may be doing both of these operations. For example Spearman, working on material given by psychological tests, was led to formulate the model of a single *g*–factor which he identified with intelligence. This was working from data to model. But once the model was given it was compared with further data. Some of these could not be made to fit and the model was modified to include further factors. This model was now compared with further observation, and so on. Most science progresses in this way and we shall find in the examples considered below that it is not always easy to classify them into component or factor analyses. This need not worry us. They are often both at different stages of the cycle from experiment to hypothesis and back again.

3.3 We will suppose, as before, that we have a matrix of observations x_{ij}, and we wish to consider whether they can arise from a situation with a structure given by

$$x_i = \sum_{k=1}^{p} a_{ik} f_k + b_i s_i + c_i \varepsilon_i \qquad i = 1 \ldots p \qquad (3.1)$$

To avoid confusion we omit the suffix j relating to the

particular observation on the ith variate. Here the f_k are factors which may appear in more than one x, s_i is a factor specific to x_i and ε_i is an error term. The f, s, and ε are regarded as having unit variance and assumed to be all independent inter se.

With this degree of generality the model is underdetermined. In fact, by a component analysis we can always express the x's in terms of f's without invoking specific or error terms at all. But we shall often consider the case where there are only one, two or three f's which is equivalent to setting up a model where a number of the coefficients a are zero. In such a case any residual variation may be ascribed to specifics or error terms.

3.1 We shall also find it necessary to draw some distinctions between types of model.

Let us consider first of all the case where the given x's comprise all the data, that is to say we are not regarding them as a sample from some larger population.

(a) We can take a model with p components in which case no specifics or errors are involved;

(b) we can take m factors ($m < p$) and consider the residual as a simple error term

$$x_i = \sum_{k=1}^{m} a_{ik} f_k + \varepsilon_i. \qquad (3.2)$$

In this form ε_i does not necessarily have unit variance. If we wish to do so we have to insert a parameter c_i to have

$$x_i = \sum_{k=1}^{m} a_{ik} f_k + c_i \varepsilon_i . \qquad (3.3)$$

Now "error" here may have two meanings. It may refer to errors of observation in the x's or it may be a synoptic way of saying that we have run together parts of the model which ought to be kept distinct

(or a mixture of both). If we wish to attribute any residual variation to x_i it makes no difference in this model whether we call it ε or s. The only distinction arising is between the situation where residuals are particular to the x's and where they may be entangled because they comprise factors which should have been separately included in the model.

To illustrate the distinction, suppose that a number of persons are subjected to tests of some kind in all of which power of comprehension, f_1, and speed, f_2, affect performance. The ith test x_i may then be regarded as a weighted combination of f_1 and f_2.

$$x_i = a_{i1} f_1 + a_{i2} f_2 .$$

But if these tests differ in nature, they may have certain individual elements; for example, if the ith test relates to arithmetic there may be an "arithmetic ability" which also affects performance, giving

$$x_i = a_{i1} f_1 + a_{i2} f_2 + b_i s_i$$

and the score on this test by the jth individual is

$$x_{ij} = a_{i1} f_{1j} + a_{i2} f_{2j} + b_i s_{ij}$$

The same individual may not always score exactly the same on the same test (if we can give it to him on more than one occasion); or the score may be subject to observational error. We then have

$$x_{ij} = a_{i1} f_{1j} + a_{i2} f_{2j} + b_i s_{ij} + c_i \varepsilon_{ij} .$$

3.5 Further distinctions arise when we regard the x's as a sample.

(c) Taking first of all the principal component model

$$x_i = \sum_{k=1}^{p} \alpha_{ik} \phi_k + \varepsilon_i \qquad (3.4)$$

we may regard the ϕ's as fixed variates, and the ε's as errors due to sampling expressing the deviation of the particular observed x's from the population values. The problem is then to estimate the α's and the ϕ's considered as parameters of the population.

(d) The same may be true if we consider $m < p$ factors

$$x_i = \sum_{k=1}^{m} \alpha_{ik} \phi_k + \varepsilon_i \qquad (3.5)$$

(e) But if we wish to consider specific factors we have a model

$$x_i = \sum_{k=1}^{m} \alpha_{ik} \phi_k + \beta_i \sigma_i + \gamma_i \varepsilon_i \qquad (3.6)$$

in which ε is an error of observation in the x's. Apart from this we regard any member x_{ij} as an observation on a variate which is composed of two parts, a linear sum of ϕ's and a specific σ. We regard σ as distributed over the population in a particular way (usually normally with zero mean and unit variance). But we may also regard ϕ in the same way. In such a case our object is to estimate α, β and γ but not ϕ.

3.6 These distinctions are subtle but important. The student who comes to them for the first time can pass over them for the time being, merely noting their existence. The importance of the distinctions will probably be clear after we have considered some examples.

3.7 Let us revert to the model of (3.1). The f's are called *common* factors, If an f occurs in all x's it is called a *general* factor. If it occurs only in certain x's it is called a *group* factor.

The factor s_i is said to be *specific* to the variate x_i. The element ε_i is called the *unreliability* factor.

On our assumption concerning the independence and unit variance of the factors we have

$$\operatorname{cov}(x_i\, x_j) = \sum_k a_{ik} a_{jk} + b_i\, b_j \operatorname{cov}(s_i, s_j) + c_i\, c_j \operatorname{cov}(\varepsilon_i, \varepsilon_j)$$

and if we also assume that the specifics and unreliability factors are independent this reduces to

$$\operatorname{cov}(x_i, x_j) = \sum a_{ik} a_{jk}, \quad i \neq j \tag{3.7}$$

$$\operatorname{var} x_i = \sum a_{ik}^2 + b_i^2 + c_i^2, \quad i = j \tag{3.8}$$

Thus the factors b and c appear only in the variances and do not affect the covariances. If we now standardize so that the variances are unity we have

$$1 = \sum_{k=1}^{m} a_{ik}^2 + b_i^2 + c_i^2 \tag{3.9}$$

The quantity c_i^2 is called in psychological terminology *unreliability*. The statistician would more often call it an error variance. The quantity b_i^2 is called the *specificity*. The quantity $\sum a_{ik}^2$ is called the *communality* and is usually written h_i^2. The complement $1-h_i^2$ is called the *uniqueness*. $h_i^2 + b_i^2$ is called the *reliability*.

3.8 In many types of scientific inquiry the unreliability would be determined from replicated experiments as an error variance. This is undoubtedly the best course when it can be followed. But in the social sciences replication may be impossible; and in psychology it is often difficult owing to the memory factor. For this reason a good deal of attention has been given, especially in psychology, to the adjustment or estimation (or guessing) of the communality element.

3.9 Reverting to the point of view we adopted in Section 1, suppose we have n observations on p variables x_i. From the

principal component viewpoint we have

$$S = \sum_i \text{var } x_i - \sum_i l_{ij}\, l_{ik} \text{ cov } (x_j, x_k)$$

and the principal component equations are

$$\sum_j l_{ij} \text{ cov } (x_j, x_k) = \lambda_i\, l_{ik} . \qquad (3.10)$$

These are in terms of covariances. If we *now* standardize by dividing them by $\{\text{var } x_j \text{ var } x_k\}^{\frac{1}{2}}$ we do *not* get

$$\sum_j l_{ij}\, r_{jk} = \lambda_i\, l_{ik}. \qquad (3.11)$$

In fact, we only get this result if we standardize at the outset. This is another aspect of the point which we mentioned in 2.7 (d).

From the factor viewpoint, if

$$x_{ij} = \sum_k a_{ik} f_{kj}$$

$$\text{cov } (x_i, x_l) = \sum_{k,m} a_{ik}\, a_{lm}\, \frac{1}{n} \sum_j f_{kj}\, f_{mj} \qquad (3.12)$$

and if we replace $\Sigma f_{kj}\, f_{mj}/n$ by its expected value

$$\delta_{km} \quad (= 1 \text{ if } k = m$$

$$= 0 \text{ if } k \neq m\ )$$

we have

$$\text{cov } (x_i, x_l) = \sum_k a_{ik}\, a_{lk}. \qquad (3.13)$$

If the model is

$$x_{ij} = \Sigma\, a_{ik} f_{kj} + \varepsilon_{ij} \qquad (3.14)$$

we have

$$\text{cov}\ (x_i, x_l) = \Sigma\, a_{ik}\, a_{lk} + \text{var}\, \varepsilon_i\ \delta_{il} \qquad (3.15$$

Now we require to estimate the coefficient a_{ij} and if we could operate on a matrix $| \Sigma\, a_{ik}\, a_{lk} |$ our ordinary methods would apply. But we have only the estimated matrix $| \Sigma\, a_{ik}\, a_{lk} + \delta_{il}\ \text{var}\ \varepsilon_i |$. This is the same as the former matrix except for the principal diagonals, where each term is increased by var ε_i. Thus we would like to have in the main diagonal, not $\Sigma\, a_{ik}^2 + \text{var}\ \varepsilon_i$ but only Σa_{ik}^2; that is to say, if we are not to bias the estimates of the a's we must remove var ε from the diagonal terms. This is equivalent to substituting communalities for unity in the diagonals of the standardized matrix.

Likewise, if we apply principal component analysis to $| \Sigma\, a_{ik}\, a_{lk} + \delta_{il}\ \text{var}\ \varepsilon_i |$ we get biassed results and need to remove the component var ε from the diagonals before carrying out the analysis.

The treatment of communalities

3.10 Where the h_i^2 are completely unknown one method of approach has been to regard them as being at choice; and in particular to assume that they are such as to minimize the number of factors. In general this seems to assume on Nature's part a much more indulgent behaviour than we have any right to expect, but it is interesting to see what happens in such cases.

The correlation matrix with h_i^2 instead of unity in the main diagonals has then p quantities at choice. The number of independent conditions for it to have rank m is (from 2.4)

$$\frac{1}{2}\ (p - m)\ (p - m + 1)$$

Thus we can reduce a matrix of rank p to one of rank m if

$$p \geqslant \frac{1}{2}\ (p - m)\ (p - m + 1) \qquad (3.16$$

or

$$p^2 - p\,(2m + 1) + m\,(m - 1) \geqslant 0$$

leading to

$$2m + 1 - \sqrt{(8m + 1)} \leqslant 2p \leqslant 2m + 1 + \sqrt{(8m + 1)} \tag{3.17}$$

For example the following are some values of p and m

p	2	3	4	5	6	8	10	15
m	1	1	2	3	3	5	6	10

Thus for $p = 2$ and 3 we can always choose the communalities so that only one component (general factor) is required; the observed correlations are of a particular kind.

3.11 *Spearman's theorem.* If there is only one general factor the correlations are given by

$$r_{ij} = a_{i1}\, a_{j1}$$

and hence

$$\frac{r_{ij}}{r_{ik}} = \text{same for all } i$$

Thus (apart from diagonal terms) the coefficients in any two rows or columns of the correlation matrix are proportional. To put it another way, for any different a, b, c, d

$$r_{ad}\, r_{cd} - r_{ac}\, r_{bd} = 0 \tag{3.18}$$

these being the so-called tetrad differences.

Spearman's theorem says that if there is one general factor the tetrad differences vanish. This, as we have just seen, is easy to prove. The converse says that if the tetrad-differences vanish there is only one factor. This is much more difficult but under certain general conditions is true. (see Camp, 1932, *Biometrika*, 24, 418.)

Estimating communalities

3.12 Psychologists give a number of recipes for guessing communalities. One is to take the largest correlation in the corresponding column of the correlation matrix; another is to take the mean of the correlations in the corresponding column — the so-called *averoid* method.

Suppose we have, by some means or other, guessed the communalities. We then perform an analysis of the data and arrive at certain factors, fewer than p, being content to neglect the rest. We can then use the coefficients occurring in these factors to estimate new communalities, iterate and proceed until the communalities converge. What this process amounts to is that we *assume* m factors and *assume* that they account for as much as possible of the variance; this determines the communalities and consequently the "error" variances. But we have not, by a mathematical manoeuvre plus assumption, estimated the error variances such as might arise in practice. We have only estimated what they would be if the number of factors is what we think it is and the error variances are minimal.

Example 3.1

Holzinger and Harman (1941, *Factor Analysis*, Chicago University Press) using data by Gosnell and Schmidt (1936, *J. Am. Statist. Ass.*, 31, 507).

Chicago was divided into 147 election areas in the 1934 presidential elections ($n = 147$). For these areas Holzinger and Harman pick out 8 variates ($p = 8$) from 17 considered by Gosnell and Schmidt, as follows

(1) % Democratic vote and Republican vote for Lewis (a Democratic candidate).

(2) Ditto for Roosevelt (also a Democratic candidate).

(3) Party vote: % that straight party votes bore to total votes.

(4) Median rental in dollars.

(5) Home ownership: % of total families owning their own houses.

(6) Unemployment: % unemployed in 1931 of gainfully employed workers over nine years in age.

(7) Mobility: % of total families living more than one year at their present address.

(8) Education: % of population over 17 years of age who completed at least ten grades at school.

Holzinger and Harman assumed a two factor pattern for this complex, estimated communalities by averaging correlations and did a centroid analysis to obtain the following coefficients

Variate	First factor	Second factor
1	.69	−.28
2	.88	−.48
3	.87	−.17
4	−.88	−.09
5	.28	.65
6	.89	.01
7	−.66	−.56
8	−.96	−.15

The first factor was found to contribute about 62% of the variance and the second about 14%, making about 76% for the two.

The absolute values of these coefficients are not of primary importance; we are more concerned with their signs and their relative magnitudes. We note that the first factor appears strongly in variates 1, 2, 3 and 6 in a positive way and negative in variates 4, 7 and 8. Those areas possessing it had more Democrats, greater unemployment, greater mobility and less education than the Republican complement. Perhaps we should emphasize that this does not prove that democratic voting is caused by the other factors. All we are doing is to explain the variation in terms of two "factors", the first being apparently a compound of the features we have listed.

The second factor is much less important but seems significant. The greatest coefficients are those for home-ownership (positive) and votes for Roosevelt and immobility (negative). This may be even more tentatively identified with a "home-permanency" factor.

I must leave it to American citizens to judge whether these results are reasonable. In any case the methodology is interesting and seems capable of development.

Example 3.2

(Rhodes, 1937, *J. R. Statist. Soc.*, **100**, 18.)

Rhodes took monthly series for the 48 months July 1931 – June 1935 of 13 series contributing to the "Economist" Index of Business Activity: e.g. Employment, Coal Consumption, Merchandise on Railways, Postal Receipts, Imports of Raw Materials, Bank Clearings. He assumed a "general business activity" factor and considered other effects as residual, giving him, in our notation

$$x_i = a_i f_i + b_i \tag{3.19}$$

He determined the a's by a least-squares solution of

$$\text{cov}\,(x_i, x_j) = a_i a_j,$$

after taking logarithms to give

$$\log a_i + \log a_j = \log \text{cov}\,(x_i, x_j). \tag{3.20}$$

The results were not very unlike those given by the "Economist" Index. Note that

(a) successive values of the series are correlated so that the observations are not independent;

(b) the least-squares solution is therefore somewhat heuristic;

(c) Rhodes calculated the residuals after extraction of the first factor and came to the conclusion that they were not independent. This led him to suggest the existence of group factors.

Rather oddly, Rhodes' pioneer work in this field does not seem to have been followed up. This may be due to the fact that economists disclaim a knowledge of multivariate methods (at least of this kind) and statisticians are nervous about the serial correlation effect in component analysis. The subject would probably repay further study.

Example 3.3

(Burt and Banks, 1941, *Ann. Eugen. Lond.*, 13, 238)

2,400 male volunteers for the Royal Air Force were divided into 8 age groups. The total range of age was 17 – 38 and the results for the 8 groups were so similar that consolidated figures could be presented and are discussed here.

An analysis was carried out on "types" of body build. The nine variates specified below were available.

		f_1	f_2	f_3
1	Standing height	+	–	–
2	Sitting height	+	–	–
3	Arm length	+	–	–
4	Leg length	+	–	+
5	Thigh length	+	–	+
6	Abdomen girth	+	+	+
7	Hip girth	+	+	–
8	Shoulder girth	+	+	–
9	Weight	+	+	–

Six factors were extracted by the centroid method. The first three accounted for 55%, 13.5% and 10.1% (78.6% in all) of the variance. The remainder were small and of very

doubtful significance.

The foregoing table gives the signs of the coefficients of the factors – we have not bothered to write down the actual values. The main factor f_1 is positively associated with every measurement. This was identified as a general "size" factor. (This might mean that there was some quality of an individual which determined the size of his body or that there are qualities which preserve some rough sort of proportionality in different individuals.)

Factor 2 has negative signs for measurements on the length of the body and positive signs for those on girth and on the weight. This was held to support the dichotomy, proposed by some anthropometrists, into two types of body build, the leptosomic or lean type and the pyknic or thickset type. If this factor has any reality (e.g. if it could be identified in a gene) it would imply the existence of some effect which might vary from one end of a range, involving a very spare lean individual, to the other end at which the individual was very thickset – irrespective of the actual size, which is determined by the first factor.

Factor 3 has positive coefficients for measurements above the waist and negative coefficients for those below the waist. This suggests a difference between trunk length and leg length.

The model conjured up by the analysis is one for which the measurements of the individual are determined by three independent factors. The first determines how big the individual is to be; the second decides whether he is to be leptosomic or pyknic or somewhere between the two; the third differentiates his trunk and leg lengths. Whether this is a plausible model must be left for the anthropometrist or the geneticist. The analysis has shown that about 80% of the variation can be accounted for by three independent components, but it is open to question whether this is more than a convenient mathematical representation.

Example 3.4

(Harper and others, 1950, *Brit. J. App. Phys.* 1, 1)

A number of measurements were made on a set of plastics with the following results :

	f_1	f_2	f_3
A. Tensile strength	+ .86	+ .10	+ .07
B. % elongation at break	+ .75	+ .03	+ .10
C. % elongation under deadload in 5 minutes	+ .92	+ .14	+ .30
D. % recovered of C in 5 minutes	+ .35	+ .08	+ .11
E. B.S. Hardness number	+ .92	+ .27	+ .29
F. % indentation in 24 hours by a loaded wire	+ .81	+ .16	+ .42
G. Minus C° to which lowered before breaking in a standard way	+ .71	− .57	− .41
H. log volume of resistivity	+ .69	+ .17	− .36
I. Dieletric constant	+ .68	+ .19	− .50
J. Power factor	+ .17	− .58	+ .41

The analysis was done by an estimation of communalities and a centroid method.

Three factors were extracted, accounting respectively for 52.2, 8.7 and 10.9% of the variance (totalling 71.8%). Note that the second accounts for less than the third. This can happen in a centroid analysis, but not, of course, in a principal component analysis.

The interpretation of these results is a matter of difficulty. With the exception of the first factor there seems nothing to suggest reality. The first factor itself seems to have something to do with molecular cohesion.

This paper has one point of theoretical interest, namely that the estimated communalities can be judged in the light of what is known about the experimental error.

Example 3.5

(Buckatzsch, 1947, *Population Studies*, 1, 229)

Buckatzsch's study concerned the influence of social conditions on mortality rates and the latter were taken as dependent variables. From his results, however, we can consider the possibility of expressing the independent variables in terms of principal components.

For 81 County Boroughs (= large towns and cities) in England and Wales in the 1931 Census the following were ascertained :

		f_1	f_2
1.	% families living in one room	+	−
2.	% males unemployed	+	−
3.	% of male population in working classes	+	+
4.	% women engaged in factory occupations	+	+
5.	Latitude N of $50^{\circ}30'$	+	+

The signs of coefficients are shown for two factors extracted, which accounted for about 80% of the variance. (Buckatzsch estimated communalities and proceeded by principal components.)

The results are very tentative. The first factor is identified with "social conditions". The second, so far as it means anything, seems associated with the employment of women in factories, the coefficient of this variate in f_2 being much larger than for the others.

4. FUNCTIONAL RELATIONSHIP

4.1 We have referred in the previous chapter to the importance of distinguishing between various kinds of model which are apt to become confused in factor analysis. It is equally important to distinguish certain types of model in other multivariate situations.

The source of a good deal of the trouble lies in the over-facile way in which a statistician gives the name "error" to any discrepancy between model and observation and then is misled by his own terminology to postulate stochastic behaviour in the "error" term. An extreme example may make the point clear.

4.2 Suppose a physicist begins with the intention of investigating the behaviour of a gas under changes in pressure and volume. By a few primitive experiments he is led to expect a relation of the Boyle type

$$\log P + \log V - \log K = 0 \qquad (4.1)$$

where P is the pressure and V the volume; but he forgets to take changes in temperature into account. He may then wish to conduct more exact experiments to verify the law more closely and takes a number of readings of the variables P and V over a period of time. If he plots the variables on a logarithmic scale he will get points lying nearly on a straight line, and the best line he can draw will give him an estimate of the constant K. The problem arises because the points do not lie exactly on a straight line.

Such a situation, simple as it is, requires at least three different techniques of analysis according to the way we approach it.

(a) We may regard (4.1) as the true underlying law of behaviour and conduct our experiment so that P is given what values we choose without error. Any discrepancy between observation and hypothesis is then due to errors in log V. These are errors of observation and the model is (for observed V, say v)

$$\log v = \log P - \log K + \varepsilon_1. \qquad (4.2)$$

We now suppose that ε_1 is a random variable or variate and have a simple regression model. Conversely we might arrange the experiment so that V was determined without error and p observed, in which case we should have a different regression model.

$$\log p = -\log V + \log K + \varepsilon_2 . \qquad (4.3)$$

It is not obvious that the same methods of estimation applied to (4.2) and (4.3) would lead to the same estimate of K.

(b) As a second model we may regard (4.1) as the true law of behaviour but contemplate errors of observation in both variables. The model now is that we observe p, v, given by

$$p = P + \varepsilon$$

$$v = V + \eta \qquad (4.4)$$

and P, V are related by (4.1). We now have a completely different situation.

(c) Again, we may suspect that (4.1) is not the right model and that we have left out a part of the true model (as in this case we have, namely the temperature T). We do not quite know what we have omitted, but we hope that it is not very important. We can now suppose, if we like, that P and V are observed

without error, but we shall have for the model

$$\log P + \log V - \log K = \zeta \qquad (4.5)$$

where ζ stands for "something-left-out-which-we-piously-hope-is-not-very-important". If we like, we can assume that this behaves like a random variable, in which case the model (4.5) becomes formally equivalent to either (4.2) or (4.3), whichever we prefer. But obviously we are making an enormous assumption here, very unlike the more customary assumption concerning errors of observation, which we know on empirical grounds to behave stochastically. The statistician makes this assumption very often, more often perhaps than he realises.

4.3 I assume that the reader is already acquainted with the standard regression theory required to deal with situations of type (4.2) where there are several independent variables. (For a more detailed discussion see the expository articles on "Regression, Structure and Functional Relationship", *Biometrika*, 1951, **38**, 11 and 1952, **39**, 96). In this chapter we shall be concerned with models of the second kind where both variables are subject to errors of observation.

4.4 We consider the case where two variables U and V are related by a linear function

$$V = \alpha + \beta U . \qquad (4.6)$$

These variables are not directly observed. In fact if d and e are random variables we observe

$$x'_U = U + d \qquad (4.7)$$

$$y'_V = V + e . \qquad (4.8)$$

Throughout we take d and e to be independent.

Substituting in (4.6) we have

$$y' = V + e = \alpha + \beta U + e$$

$$= \alpha + \beta(x' - d) + e$$

$$= \alpha + \beta x' - \beta d + e \qquad (4.9)$$

Now for this to be treated as a regression of y' on x' we should require the "residual" $(-\beta d + e)$ to be distributed independently of x'. But from (4.7) x' and d are not independent and ordinary regression theory does not apply.

4.5 We shall discuss four cases :

(1) Berkson's case : the residual $(-\beta d + e)$ is independent of x'.

(2) Geary's case : U is a random variable but is not normal.

(3) The classical case 1 : U is not a random variable.

(4) The classical case 2 : U is a random normal variable.

It appears that the problem of estimating α and β is soluble in cases (1) and (2) but not in (3) and (4) without some further restriction among the conditions of the problem.

Berkson's case

4.6 It is convenient to consider Berkson's case first because it can be referred back to the regression model. Beginning effectively from (4.9) Berkson considers a situation where d is independent of x' by introducing the idea of a controlled variable. (If x' is a 'fixed' variable we mean that d is mathematically independent; if x' is a variate d is statistically independent.)

Consider an experiment in which we are testing Boyle's law by adjusting pressures and then measuring the corresponding volumes. We decide, shall we say, to adjust the pressure to 28 lbs per square inch and do so *as far as our apparatus and instruments will allow*. We shall commit an error in calling

this 28 lbs per square inch but we ignore this error and proceed as if 28 is the correct value. We require to assume only that positive and negative errors are equally likely. This curious and subtle manoeuvre alters the model of (4.9). In fact if we call this observed and controlled variable X the 'true' value, u is a random variable connected with X by the relation

$$u = X + d$$

or better perhaps

$$X = u - d. \tag{4.10}$$

The linear relation (4.6), with (4.8) and (4.10) then becomes

$$\begin{aligned} y' &= \alpha + \beta(X + d) + e \\ &= \alpha + \beta X + \beta d + e. \end{aligned} \tag{4.11}$$

Now this is an ordinary regression model with a 'fixed' variable X and a residual $\beta d + e$ which is independent of X. The standard theory applies and we have unbiased estimators of α and β given by

$$a = \bar{y}' \tag{4.12}$$

$$b = \frac{\Sigma y' X}{\Sigma X^2} \tag{4.13}$$

where X is measured about its mean. The usual t-test for significance applies under assumptions of normality; we shall see that this ceases to be true in non-linear cases.

4.7 Those who refer to Berkson's paper (1950, *J. Am. Statist. Assoc.*, **45**, 164) should also consult Lindley (1953, *Biometrika*, **40**, 47). Berkson was ostensibly discussing the question whether there are two regression lines in a bivariate situation. There certainly are, and Berkson's negative answer really relates to the problem of functional relationship.

Non-linear Berkson case (Geary's extension)

4.8 Geary (1953, *J. Am. Statist. Ass.* **48**, 94) has extended Berkson's argument to the non-linear case.

Suppose that X is observed at m points and, since the scale is at choice, we arrange it so that odd moments of the X's vanish, i.e.

$$\sum_{i=1}^{m} X^{2k-1} = 0, \quad k = 1, 2, \ldots. \tag{4.14}$$

Then if E relates to expectations over repetitions of the experiment *with the same* X's we find, from (4.11)

$$\sum_{i=1}^{m} X_i^{2k} E(y_i) = \alpha \sum_{i=1}^{m} X_i^{2k}$$

$$= m \alpha \mu_{2k}, \quad k = 0, 1, \ldots \tag{4.15}$$

$$\Sigma X_i^{2k-1} E(y_i) = \beta \Sigma X_i^{2k}$$

$$= m \beta \mu_{2k}, \; k = 1, 2, \ldots \tag{4.16}$$

The simplest solution is given by $k = 0$ in (4.15) and $k = 1$ in (4.16), leading to

$$m\alpha = \Sigma E(y_i) \tag{4.17}$$

$$m\beta\mu_2 = \Sigma X_i E(y_i), \tag{4.18}$$

and consistent estimators of α and β are then

$$a = \frac{1}{m}\Sigma y \tag{4.19}$$

$$b = \frac{1}{m\mu_2}\Sigma xy. \tag{4.20}$$

These are also maximum likelihood estimators if the errors d and e are normal and in any case (for zero means in d and e) are Gauss-Markoff estimators. This confirms our previous analysis of Berkson's case.

4.9 Consider now the case when the underlying relationship is non-linear, say, cubic of the form

$$\eta = \alpha + \beta u + \gamma u^2 + \delta u^3 . \tag{4.21}$$

We standardize as in (4.10) and then have

$$y' = \alpha + \beta(X - d) + \gamma(X - d)^2 + \delta(X - d)^3 + e \tag{4.22}$$

Multiplying by 1, X, X^2, X^3 and summing over $i = 1 \ldots m$ we have

$$\nu_{01} = \zeta + \gamma \mu_2 \tag{4.23}$$

$$\nu_{11} = \eta \mu_2 + \delta \mu_4 \tag{4.24}$$

$$\nu_{21} = \zeta \mu_2 + \gamma \mu_4 \tag{4.25}$$

$$\nu_{31} = \eta \mu_4 + \delta \mu_6 \tag{4.26}$$

where

$$\nu_{k1} = \frac{1}{m} \Sigma X^k E(y_i) \tag{4.27}$$

$$\zeta = \alpha + \gamma \operatorname{var} d \tag{4.28}$$

$$\eta = \beta + 3\delta \operatorname{var} d \tag{4.29}$$

We then find for γ and δ

$$(\mu_4 - \mu_2^2)\gamma = \nu_{21} - \mu_2 \nu_{01} \tag{4.30}$$

$$(\mu_6 \mu_2 - \mu_4^2)\delta = \mu_2 \nu_{31} - \mu_4 \nu_{11} \tag{4.31}$$

This will give us consistent estimators for γ and δ in terms of the observables.

4.10 But there are two remarkable features about this situation. The first is that, although we can estimate γ and δ, we cannot estimate α and β. The quantities which occur in equations like (4.23) – (4.26), however many of them we take, are ζ and η, which we can accordingly estimate. But from (4.28) and (4.29) we see that we cannot estimate α and β without a knowledge of the nuisance parameter var d. The only way round this difficulty so far suggested (by Geary himself) is to replicate the experiment. Since the whole analysis is applicable mainly to experimental situations this is usually possible.

4.11 The second feature is that no ordinary test of significance is applicable. In fact the "error variance" is no longer $(-\beta d + e)$ but varies from one value of X to another. The usual tests dependent on variance-ratios such as Student's t do not apply. The problem of testing significance in this case has not, apparently, been solved.

Geary's case

4.12 In the case we now consider the random variables u and v are connected by

$$v = \alpha + \beta u \tag{4.32}$$

and there are errors of observation in u and v :

$$x' = u + d \tag{4.33}$$

$$y' = v + e \tag{4.34}$$

We may attach a dummy variate to α to make the equation homogeneous and write the more general form in p variates as

$$\Sigma \beta_i u_i = 0 \tag{4.35}$$

with

$$x_i' = u_i + d_i$$

If parameter θ_i relates to x_i' we have that the cumulant generating function of the x_i is the sum of the c.g.f.'s of u_i and d_i and on expansion

$$\Sigma \kappa_{p_1\ p_2\ \dots\ p_r}(x') \frac{\theta_1^{p_1}\ \theta_2^{p_2}\ \dots\ \theta_r^{p_r}}{p_1!\ p_2!\ \dots\ p_r!} = \text{the sum of similar expressions in the cumulants of } u_i \text{ and } d_i.$$

On the right the terms involving the cumulants of d do not contain any product terms, for all the d's are independent, and only powers of individual θ's can appear. Thus we have

$$\kappa_{p_1\ p_2\ \dots\ p_r}(x') = \kappa_{p_1\ p_2\ \dots\ p_r}(u) \qquad (4.36)$$

so long as there are at least two p's involved. Thus we can take cumulants calculated from observation as estimators of the cumulants of the u's.

Now from (4.35) we see that the characteristic function of $\Sigma\beta_i u_i$ is unity and thus

$$1 = \int \text{Exp}\,(\theta\Sigma\beta_i u_i)\, dF(u_1) \dots dF(u_p)$$

$$= \phi(\theta\beta_1, \theta\beta_2, \dots \theta\beta_p), \text{ say.} \qquad (4.38)$$

From the property of homogeneous functions we then have

$$\Sigma\beta_i \frac{\partial \log \phi}{\partial \beta_i} = 0$$

giving

$$\Sigma\beta_i\, \kappa_{p_1, p_2, \dots p_{(i-1)}, p_i+1, p_{i+1} \dots p_r} = 0 \qquad (4.39)$$

This gives us a set of linear equations in the β's and the κ's which, in virtue of (4.38), are calculable from the observations. Hence we may estimate the β.

Unfortunately, this method fails to give a result in the

case when we most need it, namely when the variation is normal; for then cumulants of order higher than two all vanish and the equations (4.39) are empty except for a few in the cumulants κ_{11} which are not enough to determine the β's.

4.13 Moreover, in cases where the distribution of the u's is nearly normal the values of κ's observed will often have a high sampling error compared with the true values, and estimates of the β's given by (4.39) may therefore be rather wide of the mark. It is somewhat strange that normality, which we often invoke to make a method work at all, should stultify this particular method. There appears, however, to be some fundamental connection between normality and indeterminacy in this and related subjects which has not been fully brought to light.

Classical case 1

4.14 We now revert to the model

$$V = \alpha + \beta U$$

where V, U are not random variables and each is subject to error:

$$x' = U + d$$

$$y' = V + e$$

If d, e are normal variates the likelihood is proportional to

$$\frac{1}{\delta_2^n} \operatorname{Exp}\left\{ -\frac{1}{2\delta_2^2} \Sigma(y' - V)^2 \right\} \frac{1}{\delta_1^n} \operatorname{Exp}\left\{ -\frac{1}{2\delta_1^2} \Sigma(x' - U)^2 \right\}$$

where δ_1^2, δ_2^2 are the variances of d, e, and are supposed the same for each observation. Thus, except for a constant,

$$\log L = -n \log \delta_1 - n \log \delta_2 - \frac{1}{2\delta_1^2} \Sigma(x' - U)^2 - \frac{1}{2\delta_2^2} \Sigma(y' - \alpha - \beta U)^2 \qquad (4.40)$$

If we now maximise for the $n + 4$ parameters δ_1, δ_2, α, β, U_i we find

$$\frac{\partial}{\partial \delta_1} = \frac{-n}{\delta_1} + \frac{2}{2\delta_1^3} \Sigma (x' - U_i)^2 = 0$$

giving

$$\hat{\delta}_1^2 = \frac{1}{n} \Sigma (x' - U_i)^2 \qquad (4.41)$$

Similarly

$$\hat{\delta}_2^2 = \frac{1}{n} \Sigma (y' - \alpha - \beta U_i)^2 \qquad (4.42)$$

Differentiating for U_i, α, β, gives us

$$\frac{1}{\delta_1^2} (x_i' - U_i) + \frac{1}{\delta_2^2} (y_i' - \alpha - \beta U_i) \beta = 0 \qquad (4.43)$$

$$\Sigma (y' - \alpha - \beta U) = 0 \qquad (4.44)$$

$$\Sigma U (y' - \alpha - \beta U) = 0 \qquad (4.45)$$

Now summing (4.43) and using (4.45) we have

$$\Sigma (x' - U) = 0 \qquad (4.46)$$

Also from (4.41) and (4.42), using (4.43), we find

$$\beta^2 \hat{\delta}_1^2 = \hat{\delta}_2^2 \qquad (4.47)$$

This is unacceptable, for we have no reason to suppose that the error variances of d and e are proportional to β^2. The basic reason for the breakdown of the maximum likelihood method seems to be that we are trying to estimate too many parameters. There are here $n + 4$ parameters against only n observations.

4.15 It appears that no progress can be made in this direction unless we make some assumption about the error variances. We shall assume that their ratio is known :

$$\delta_1^2 / \delta_2^2 = \lambda \qquad (4.48)$$

The maximum likelihood equations are now

$$\log L = \text{constant} - \frac{2n}{\delta_1} + \frac{1}{\delta_1^3} \Sigma(x' - U)^2 + \frac{\lambda}{\delta_1^3} \Sigma(y' - \alpha - \beta U)^2 \qquad (4.49)$$

$$x_i - U_i + \lambda \beta (y_i' - \alpha - \beta U_i) = 0 \qquad (4.50)$$

$$\Sigma (y' - \alpha - \beta U) = 0 \qquad (4.51)$$

$$\Sigma U (y' - \alpha - \beta U) = 0 \qquad (4.52)$$

Without loss of generality take $\bar{x} = 0$, $\bar{y} = 0$. Then from (4.51)

$$\alpha + \beta \bar{U} = 0$$

and from (4.50)

$$-\bar{U} + \lambda \beta (\alpha + \beta \bar{U}) = 0.$$

Hence

$$\bar{U} = 0 \qquad (4.53)$$

$$\hat{\alpha} = 0 \qquad (4.54)$$

and for the estimator of β from (4.50) we find

$$U_i = \frac{x_i' + \lambda \beta y_i'}{1 + \lambda \beta^2} \qquad (4.55)$$

and hence

$$\Sigma (y' - \beta x') (x' + \lambda \beta y') = 0$$

$$\lambda \hat{\beta} \Sigma y'^2 + (1 - \lambda \hat{\beta}^2) \Sigma x' y' - \hat{\beta} \Sigma x'^2 = 0 \qquad (4.56)$$

This gives us a formula for the estimation of β

$$\hat{\beta}^2 \lambda \Sigma x'y' - \beta (\lambda \Sigma y'^2 - \Sigma x'^2) - \Sigma x'y' = 0 \qquad (4.57)$$

If

$$\theta = \frac{\Sigma y'^2 - \frac{1}{\lambda} \Sigma x'^2}{2 \Sigma x'y'} \qquad (4.58)$$

then

$$\hat{\beta} = \theta + \sqrt{(\frac{1}{\lambda} + \theta^2)} \qquad (4.59)$$

where we take the positive root to maximise the likelihood.

4.16 We notice that if $\lambda = 0$ (corresponding to no error in U) the estimator of β from (4.56) is

$$\hat{\beta} = \frac{\Sigma x'y'}{\Sigma x'^2}$$

and, in fact, we are back at a regression situation. Similarly, if λ is infinite we have a regression of U on V. For values of λ between 0 and infinity the line $V - \alpha - \beta U = 0$ lies between these two extremes. In particular if $\lambda = 1$ the line takes the direction of the first principal component. It was this line which Frisch called "diagonal regression".

4.17 If U and V are measured about their means the estimator of $\hat{\alpha}$ is zero and

$$\Sigma V^2 = \beta \Sigma UV = \beta^2 \Sigma U^2 \qquad (4.60)$$

Thus

$$E(\Sigma x'^2) = \Sigma U^2 + \frac{n-1}{n} \delta_1^2$$

$$E(\Sigma x'y') = \Sigma UV$$

$$E(\Sigma y'^2) = \Sigma V^2 + \frac{n-1}{n} \frac{\delta_1^2}{\lambda}$$

Hence θ of (4.58) converges in probability to

$$\frac{\Sigma V^2 + \frac{n-1}{n}\frac{\delta_1^2}{\lambda} - \frac{1}{\lambda}\left(\Sigma U^2 + \frac{n-1}{n}\delta_1^2\right)}{2\Sigma UV} = \frac{\beta^2 - \frac{1}{\lambda}}{2\beta},$$

and $\hat{\beta}$ from (4.59) then becomes

$$\frac{\beta^2 - \frac{1}{\lambda}}{2\beta} + \sqrt{\left\{\left(\frac{\beta^2 - \frac{1}{\lambda}}{2\beta}\right)^2 + \frac{1}{\lambda}\right\}} = \beta \qquad (4.61)$$

Thus $\hat{\beta}$ is a consistent estimator of β.

4.18 But even here there are still peculiarities in the situation. Consider what happens if we estimate δ_1^2. From (4.49) we have

$$\hat{\delta}_1^2 = \frac{1}{2n}\left\{\Sigma (x' - U)^2 + \lambda \Sigma (y - \hat{\alpha} - \hat{\beta} U)^2\right\}$$

with $\hat{\alpha} = 0$ and using (4.50) we have

$$\hat{\delta}_1^2 = \frac{1}{2n}\left(1 + \frac{\lambda}{\lambda^2\beta^2}\right) \Sigma (x' - U)^2. \qquad (4.62)$$

Using (4.50) again we have

$$\begin{aligned}
\hat{\delta}_1^2 &= \frac{\lambda}{2n}(1 + \lambda\beta^2) \Sigma (y' - \hat{\beta} U)^2 \\
&= \frac{\lambda}{2n}(1 + \lambda\beta^2) \Sigma y' (y' - \hat{\beta} U) \\
&= \frac{\lambda}{2n}(1 + \lambda\beta^2) \Sigma y' \left\{y' - \frac{\hat{\beta}(x' + \lambda\beta y')}{1 + \lambda\beta^2}\right\} \\
&= \frac{\lambda}{2n} \Sigma y' (y' - \hat{\beta} x')
\end{aligned} \qquad (4.63)$$

Hence

$$\frac{2n\delta_1^2}{\lambda} = \Sigma y'^2 - \hat{\beta}\Sigma x'y'$$

and

$$\hat{\delta}_1^2 = \frac{1}{2}\lambda\{\frac{1}{2n}\Sigma y'^2 + \frac{1}{2\lambda n}\Sigma x'^2 - \frac{1}{2}[(\frac{1}{n}\Sigma y'^2 - \frac{1}{\lambda n}\Sigma x'^2)^2 + \frac{4}{\lambda n}\Sigma x'y']^{\frac{1}{2}}\}$$

$$\rightarrow \frac{1}{2}\lambda\{\frac{1}{n}\Sigma U^2 + \frac{1}{\lambda}\delta_1^2 - \frac{\beta}{n}\Sigma UV\}$$

$$\rightarrow \frac{1}{2}\delta_1^2 \qquad (4.64)$$

and thus $\hat{\delta}^2$ is not a consistent estimator of δ_1^2. We need to multiply it by 2 to get such an estimator.

4.19 A more comprehensive treatment of the classical case 1 has been given by Lindley (1947, *Supp. J. Roy. Statist. Soc.* 9, 218), who also shows that if λ is known the same equations of estimation arise when u and v are jointly normally distributed – our so-called Case 2. We shall not pursue the matter further here except to point out that unbiased estimators of functional parameters may be obtained in certain cases by distribution-free methods.

4.20 Suppose that we have a set of observations x', y' and that they can be divided into two groups such that the values of v for one group are all greater than the values of v for the other. (This is not the same as dividing them according to the values of y', which of course is a trivial matter). We can, to put it broadly, then determine a typical point, such as a centre of gravity, for each group and make our linear relation pass through the two points. A fortiori, if we can divide into more than two groups we can fit a relationship of greater accuracy; the extreme case (Theil, 1950, *Indagationes Mathematicae*, 12, No. 2) occurs when the order of the points with regard to one variable, say y', is the same as the order

with regard to the underlying v, in which case we can determine a distribution-free regression line.

4.21 The assumption about division into groups can often be made when the errors of observation are small compared to the intervals between the variables u and v. Logically, the subject offers more of a problem. Wald (1940, *Ann. Math. Statist.* 11, 284), who first discussed the division into two groups, gave a condition for the validity of the procedure but unfortunately it is not obeyed by normal variables u and v.

4.22 We shall have to leave the subject at this point, but it is to be noted that many important questions arising in connection with the analysis of functional relationships remain unsolved. For some work in this field reference may be made to Neyman, J. and Scott, E.L. (1948) *Econometrica*, **16**, 1 and to Creasy, M.A. (1956), *Jour. Roy. Statist. Soc.*, B, **18**, 65. See also Brown, R. L. (1957), *Biometrika*, **44**, 84.

5. CANONICAL ANALYSIS

5.1 Let us recapitulate some of the basic results of ordinary regression theory. We suppose that we have a number of observations on each of p variables $X_1 \ldots X_p$ which are not necessarily random variables (or, in our terminology, variates). There are certain unknown constants $\beta_0, \beta_1, \ldots \beta_p$ which it is our object to estimate. The model we consider is that an observed variate, y, is given by

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \qquad (5.1)$$

where ε is a random variable. Without loss of generality we may suppose it to have zero mean (for any constant can be absorbed into β_0). For expository simplicity we can take a dummy variable X_0 attached to β_0, having always the value unity and so write

$$y = \sum_{j=0}^{p} \beta_j X_j + \varepsilon \qquad (5.2)$$

We also assume as part of the model that ε has the same variance for any set of X's whatsoever. The X's may be functionally dependent, so that this model includes the case of a curvilinear regression line.

5.2 The solution to the problem of estimation is well known. Without appeal to normality we can derive unbiassed estimators of minimal variance from the Gauss-Markoff theorem in least squares; and if in addition we postulate normality the estimators are maximum likelihood estimators.

5.3 The principal point to notice about this model is that it is not multivariate at all. It is *multivariable* in the sense that there are p variables X, but they are not random variables.

The only variates in the situation are ε and the equivalent y. Thus multiple regression is not multivariate. We could make it so by relaxing our restrictions on ε so that, for example, different X's gave different variances to ε, or that successive values of ε were not independent. But the simple standard situation is univariate.

5.4 A wide class of problems in statistics can be expressed in regression form. Apart from ordinary regression theory itself we may, for instance, take dummy variables for our X's, allowing them to be unity if the observation falls into a given class and zero elsewhere, to get the customary models of variance analysis in experimental design.

5.5 What has caused some confusion in the past is that since the X's can be chosen how we like, they may in particular be the values of random variables. If we choose numbers at random from a $(p + 1)$ multivariate population we can consider relations of type

$$y = \sum_{j=0}^{p} \beta_j x_j + \varepsilon \qquad (5.3)$$

which looks like a linear functional relation between y, the x's and ε. It is, in fact, a relation of the functional kind, but it is still asymmetric in the sense that the x's are supposed free from error whereas y is subject to the error ε. For this reason we get a different regression line if we pick out one of the other variates x as the independent variate subject to error. The model is distinct from

$$y - \Sigma \beta_j x_j = \varepsilon \qquad (5.4)$$

where y and the x's are free from error and ε is an unobservable residual term expressing imperfection in the model; and from

$$y - \Sigma \beta_j x_j - \varepsilon = 0 \qquad (5.5)$$

where y, x and ε are variates free from error and ε is now observable. We have made this distinction before but it will

do no harm to make it again.

5.6 When the X's are powers of a single variable X it is possible to simplify the exposition and analysis of a regression model by choosing our variables to be orthogonal polynomials. (The ultimate results, of course, are the same.) This is particularly useful when the values of X are equidistant, in which case many of the quantities required can be (and have been) tabulated once and for all.

It is equally possible to "orthogonalize" a regression situation for any arbitrary set of variables X. This possibility does not seem to have been much discussed in the literature. But it throws some new light on certain old but unsolved problems; particularly (a) how many variables do we take ? (b) how do we discard the unimportant ones ? and (c) how do we get rid of multicollinearities in them ?

5.7 In fact, let us suppose that we do a principal component analysis on the variables $X_1 \ldots X_p$ and arrive at new components $\zeta_1 \ldots \zeta_p$. These will give us

$$y = \sum_{j=0}^{p} \alpha_j \zeta_j + \varepsilon, \qquad (5.6)$$

where the α's are linear functions of the β's. On our model the Gauss-Markoff theorem applies to give us estimators of the α's which are the same linear functions of the estimators of the β's. We therefore lose nothing by the transformation, except the time spent on the arithmetical labour of finding it. But our ζ's are now all orthogonal and we have at once

$$\alpha_j = \frac{\Sigma y \zeta_j}{\Sigma \zeta_j^2}, \qquad (5.7)$$

the summation extending over the sample; and this is easily converted into sums of type Σyx, which are already known. Further, the reduction in variance due to the fitting of ζ_j is $\alpha_j^2 \lambda_j$ where λ_j is the characteristic root corresponding to ζ_j. We can thus see how important each contribution is and decide whether any ζ's can be discarded as unimportant.

This in turn will allow us to see how far the X's are important.

We can also apply the "Student" t-test to the "significance" of the α's; and this will also apply even when the original x's are variate values chosen by a random process.

Example 5.1

(Stone, J.R.N., 1945, *J. Roy. Statist. Soc.*, **108**, 308.)

In some studies of demand analysis Stone considered the consumption of beer in the United Kingdom for the years 1920–1938 inclusive. He was interested in an equation of type

$$\log q = a + b \log \mathrm{Q} + c \log p + d \log \pi + f \log g + r\, t \log e \qquad (5.8)$$

where

q = consumption (bulk barrels)

Q = real income

p = retail price

π = cost-of-living index

g = gravity of the beer

t = time

e = the constant 2.718 ...

The constants a, b, c, d, f, r are under estimate. Calling the logs of Q, p, π, g, t respectively X_1, X_2 ... X_5 we have the following correlation matrix (Stone's figures):

$$\begin{bmatrix} 1 & -.610375 & -.660691 & -.507697 & .918651 \\ & 1. & .447714 & -.256291 & -.462810 \\ & & 1. & .397888 & -.831054 \\ & & & 1. & -.649439 \\ & & & & 1. \end{bmatrix}$$

A principal component analysis on this matrix gives the following

λ	3.2470	1.2753	.3859	.0700	.0218
X_1	−.5215	−.0711	.4730	.5219	−.4757
X_2	.3121	.7090	−.2161	.5942	−.0112
X_3	.4753	.0381	.8246	.0103	.3050
X_4	.3245	−.6943	−.2224	.5801	.1629
X_5	−.5471	.0935	.0105	.1945	.8087

Here, for example, $\zeta_1 = -.5215\ X_1 + .3121\ X_2 + .4573\ X_3 + .3245\ X_4 - .5471\ X_5$. If $y = \log q$ we also have for the correlations with the X's respectively −.457,536, .031,719, .899,223, .601,129, −.710,155. We will take the variables as standardized so that these correlations are also their covariances.

From the small size of λ_4 and λ_5 we seem safe in neglecting the contributions from these sources. We then have

$$\alpha_1 = \frac{\Sigma y \zeta_1}{\lambda_1} = \frac{1}{3.2470}(-.5215)(-.4575) + (.3121)(.0317)$$

$$+ (.4753)(.8992) + (.3245)(.6011) - (.5471)(-.7102)$$

$$= .3879$$

Likewise

$$\alpha_2 = -.3093$$

$$\alpha_3 = .9772$$

On our scale the variance of y is unity and the contribution of the term in ζ_i is $\alpha_i \Sigma y \zeta_i = \lambda_i \alpha_i^2$. For our first three factors these contributions are .4885, .1220 and .3685, totalling .9790; they account for about 98 per cent of the variance.

We may now go back, if we wish, to express the regression equation in terms of the *standardized* variates X. We have

$$\begin{aligned}\log q \text{ (about its mean)} &= .3879\ (-.5215\ X_1 + .3121\ X_2 + .4753\ X_3 \\ &\quad .3245\ X_4 - .5471\ X_5) - .3093\ (-.0711\ X_1 + \\ &\quad .7090\ X_2 - .0381\ X_3 - .6943\ X_4 + \\ &\quad .0935\ X_5) + .9772\ (.4730\ X_1 - .2161 \\ &\quad X_2 + .8246\ X_3 - .2224\ X_4 + .0105\ X_5) \\ &= .2819\ X_1 - .3094\ X_2 + 1.0020\ X_3 + .1233\ X_4 - .2309\ X_5\end{aligned} \tag{5.8}$$

This equation seems to make sense. The consumption is negatively related to X_5, the time variable, and over the period the consumption of beer appeared to be declining (other things being equal). The correlation with X_2, the retail price, is also negative as we should expect. On the other hand, the dependence on X_1 (real income) is positive, as also is that on the gravity (X_4). The only surprising feature is the magnitude of the coefficient of X_3, the cost-of-living index. Seeing that the effect of real income and retail price is accounted for elsewhere, one must suppose that when all prices are rising the consumption of beer goes up proportionately; but this may only be a reflection of the fact that when prices rise wages rise and more is spent on beer even if real income is constant.

It is interesting to compare this analysis with one carried out by Stone (l.c. p. 314 et seq.) by confluence methods. Stone worked out 960 regression slopes and graphed 240 bunch maps. He concludes, *inter alia*,

(1) the two most important determinants are the two price factors;

(2) the influence of income is negligible;

(3) there seems to have been little variation attributable to factors not introduced explicitly;

(4) the influence of the strength of beer is uncertain but positive.

Our analysis would support (1), (3) and (4) but is not in agreement with (2).

The real lesson to be learnt from this example, however, is that when there are collinearities in the independent variables (i.e. when some of the characteristic roots λ are zero or nearly so) no reliance whatever can be put on individual coefficients in regression equations embodying all the variables.

In fact if the observed regression is

$$y = \Sigma b_j X_j \qquad (5.9)$$

and there exists a linear relation $\Sigma l_j X_j = 0$, we can substitute for any of the X's in the regression and obtain quite different-looking results. In our present example two λ's are near zero and hence two of the five "independent" variables are nearly expressible in terms of the other three.

Suppose for example, that we omit X_1, X_2 and find the regression of y on X_3, X_4 and X_5. We get

$$y = 1.2506\, X_3 + 0.5487\, X_4 + 0.6855\, X_5. \qquad (5.10)$$

If we omit X_4 and X_5 we get

$$y = -0.021{,}25\, X_1 - 0.4722\, X_2 + 1.0966\, X_3 \qquad (5.11)$$

The equations (5.8), (5.10) and (5.11) appear very different; but *as regression equations* they are almost equivalent. The squares of the multiple correlation coefficient for the three respectively are 0.9896, 0.9676, 0.9808, indicating little variation in efficiency of prediction. (It is true that the complements of these quantities, 0.0104, 0.0324, 0.0192, are substantially different and indicate that the residual errors are by no means the same; but where all are so small this is not a very material point.)

5.8 We shall not dwell on the use of component analysis in standard regression theory, although there seems to be a fruitful field of inquiry here. The main purpose of introducing the topic here was to provide an introduction to a more general technique known as canonical analysis.

We consider a case where, instead of one variate y dependent on a set of x's we have a set of y's and a set of x's which are mutually dependent. We are not particularly interested in the relations of the x's among themselves or of the y's among themselves, but in the relationship between the two groups.

5.9 There arises here the same sort of distinction that we drew between correlation and regression, or between interdependence and dependence. From one point of view the relation of y's and x's may be considered as a symmetrical one of correlation; from the other, one set, say the y's, are regarded as dependent on the other and the x's may, as in the univariate case, be "fixed" variables. From most viewpoints it is simpler to follow the latter approach, just as it is simpler to deal with a regression model than with a correlation model. But both are in use.

5.10 Following the line of thought given earlier in this chapter we might perform a component analysis on both the y's and the x's and then investigate the relationship of the transformed variables. There is something to be said for this procedure but in canonical analysis we take a somewhat different line: we still transform x's and y's to new variables which are orthogonal (independent) but not so as to maximize the variance in any particular direction. Instead we maximize the covariances (or rather correlations) between certain members of the two sets while reducing the others to zero. Thus the relationship between the two groups is reduced to its simplest form.

Changing the notation slightly, let one set consist of p variates $x_1 \ldots x_p$ and the other of q other variates $x_{p+1} \ldots x_{p+q}$. We assume $p \leqslant q$. (In the contrary case we invert the roles of the two sets.)

Using Greek dummy suffixes for variates in the p-group and Roman suffixes for those in the q-group, take new variates defined by

$$\zeta_\alpha = \sum_\beta c_{\alpha\beta} x_\beta, \quad \alpha = 1, 2, \ldots p \tag{5.12}$$

$$\eta_a = \sum_b d_{ab} x_b, \quad a = 1, 2, \ldots q \tag{5.13}$$

If they have unit variance we have

$$\Sigma\, c_{\alpha\beta}\, c_{\alpha\gamma}\, v_{\beta\gamma} = 1 \tag{5.14}$$

$$\Sigma\, d_{ab}\, d_{ac}\, v_{bc} = 1 \tag{5.15}$$

when the v's are covariances. We now lay down the condition that the covariance of a pair from the two groups is stationary, e.g.

$$\Sigma\, c_{\alpha\beta}\, d_{ab}\, v_{\beta b} = \text{stationary} = R, \text{ say} \tag{5.16}$$

If λ and μ are undetermined multipliers, the stationary values of (5.16) subject to (5.14) and (5.15) are given by

$$\Sigma\, d_{ab}\, v_{\beta b} - \lambda\, \Sigma\, c_{a\gamma}\, v_{\beta\gamma} = 0 \tag{5.17}$$

$$\Sigma\, c_{\alpha\beta}\, v_{\beta b} - \mu\, \Sigma\, d_{ac}\, v_{bc} = 0 \tag{5.18}$$

Multiplying these two respectively by d_{ab}, $c_{\alpha\beta}$ and summing for a or α we find

$$\lambda = \mu = R$$

(5.17) and (5.18) are then soluble in their determinant vanishes, which leads to

$$\begin{vmatrix} -\lambda v_{\alpha\beta} & v_{\alpha j} \\ v_{i\beta} & -\lambda v_{ij} \end{vmatrix} = 0, \qquad \begin{matrix} \alpha, \beta = 1 \ldots p \\ i, j = 1 \ldots q \end{matrix} \tag{5.19}$$

which is equal to

$$(-\lambda)^{q-p} \begin{vmatrix} \lambda^2 v_{\alpha\beta} & v_{\alpha j} \\ v_{i\beta} & v_{ij} \end{vmatrix} = 0, \tag{5.20}$$

Premultiplying this by

$$\begin{vmatrix} \delta_{\alpha\gamma} & -\Sigma v^{ik} v_{\gamma k} \\ 0 & v^{ki} \end{vmatrix}$$

where v^{ik} is inverse to v_{ik} and v^{li} to v_{li}, we find

$$\begin{vmatrix} (-\lambda)^{q-p} & \lambda^2 v_{\beta\gamma} - \Sigma v_{i\beta} v^{ik} v_{\gamma k} & 0 \\ & \Sigma v^{li} v_{i\beta} & \Sigma v^{li} v_{ij} \end{vmatrix}$$

$$= (-\lambda)^{q-p} \left| \lambda^2 v_{\beta\gamma} - \Sigma v_{i\beta} v^{ik} v_{\gamma k} \right| = 0 \tag{5.21}$$

a determinant of p rows and columns.

We shall suppose that the roots in λ^2 are distinct. (The contrary case is tractable but of secondary interest.) We choose the p positive roots in λ. These are correlations from (5.21) and the corresponding transformations to ζ's and η's exist. We may also show that the variables have the following properties:

(1) All ζ's and η's have zero mean and unit variance;

(2) Any ζ is independent of any other ζ and any η is independent of any other η;

(3) The correlation between any ζ and any η is zero except for p correlations $\rho_1, \ldots, \rho_p$ which may be taken as the correlations between ζ_1 and η_1, ζ_2 and η_2, etc.

The dispersion matrix then becomes

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & \rho_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \rho_2 & \cdots & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & \rho_p & \cdots & 0 \\ \rho_1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 & 0 & 1 & \cdots & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \rho_p & \cdot & \cdot & \cdots & 1 & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \tag{5.22}$$

The determinant is

$$(1-\rho_1^2)\ (1-\rho_2^2)\ \cdots\ (1-\rho_p^2). \tag{5.23}$$

5.11 We notice the close resemblance between (5.21) and the characteristic determinant of component analysis. A numerical solution may proceed by iteration in the manner described in Chapter 1. A numerical example of Hotelling's will make the arithmetic clear.

Example 5.2

140 seventh-grade schoolchildren received the four tests on (a) reading speed, (b) reading power, (c) arithmetic speed and (d) arithmetic power. The correlations in performance were as follows. (Figures to four places for greater accuracy.)

$$\begin{bmatrix} 1 & .6328 & .2412 & .0586 \\ & 1. & -.0553 & .0655 \\ & & 1. & .4248 \\ & & & 1. \end{bmatrix} \qquad (5.24)$$

The determinant (5.19) is written down as

$$\begin{vmatrix} -\lambda & -.6328\lambda & .2412 & .0586 \\ -.6328\lambda & -\lambda & -.0553 & .0655 \\ .2412 & -.0553 & -\lambda & -.4248\lambda \\ .0586 & .0655 & -.4248\lambda & -\lambda \end{vmatrix} = 0 \qquad (5.25)$$

For determinants of larger order a systematic method of solution is desirable. Writing (5.25) in the form of (5.20)

$$\begin{vmatrix} \lambda^2 & .6328\lambda^2 & .2412 & .0586 \\ .6328\lambda^2 & \lambda^2 & -.0553 & .0655 \\ .2412 & -.0553 & 1.0000 & .4248 \\ .0586 & .0655 & .4248 & 1.0000 \end{vmatrix} = 0$$

"Sweep out" the λ^2 from the first row, second column and then from the second row, first column, to get

$$\begin{vmatrix} \lambda^2 & 0 & .241,200 & .058,600 \\ 0 & .599,564\lambda^2 & -.207,931 & .028,418 \\ .241,200 & -.207,931 & 1.000,000 & .424,800 \\ .058,600 & .028,418 & .424,800 & 1.000,000 \end{vmatrix} = 0$$

Then sweep out the first three items in the bottom row to get

$$\begin{vmatrix} \lambda^2 - .003,434 & -.001,665 & .216,307 \\ -.001,665 & .599,564\lambda^2 - .000,800 & -.220,003 \\ .216,307 & -.220,003 & .819,545 \end{vmatrix} = 0$$

Then sweep out the first two items in the bottom row to get

$$\begin{vmatrix} \lambda^2 - .060,525 & .056,401 \\ .056,401 & .599,564\lambda^2 - .059,867 \end{vmatrix} = 0 \qquad (5.26)$$

This is a general procedure and we end up with a determinant of type (5.26) from which the λ^2 can be obtained by the iterative method we used for principal components in Chapter 1. In our present case it is simpler to evaluate (5.26) as a quadratic in λ^2,

$$.599,564\lambda^4 - .096,156\lambda^2 + .000,442,377 = 0$$

giving $\lambda = .3945$ or $.0689$.

An examination of these values suggests that the correlation represented by 0.0688 is negligible and the relationship between the reading tests and arithmetic tests is concentrated in the first correlation. If necessary we can find the canonical variates to which this correlation corresponds. From (5.17) we have, writing c_j $(j = 1,2)$ and d_j for d_{1j}

$$\begin{aligned} & c_1 + .6328\, c_2 - .6114\, d_1 - .1485\, d_2 = 0 \\ + .6328\, & c_1 + c_2 + .1402\, d_1 - .1660\, d_2 = 0 \\ - .6114\, & c_1 + .1402\, c_2 + d_1 + .4248\, d_2 = 0 \\ - .1485\, & c_1 - .1660\, c_2 + .4248\, d_1 + d_2 = 0 \end{aligned}$$

Only three of these are independent and we solve for the ratios

$$c_1 : c_2 : d_1 : d_2 = -2.7772 : 2.2655 : -2.4404 : 1$$

Thus our new variate ζ_1 is given by a multiple of $-2.7722\, x_1 + 2.2655\, x_2$ and η_1 by a multiple of $-2.4404\, x_3 + x_4$. These are independent of ζ_2 and η_2 and the correlation between ζ_1 and η_1 is 0.3945. That between ζ_1 and η_2, ζ_2 and η_1 is exactly zero and between ζ_2 and η_2 nearly so. The relationship

is thus nearly summarised in the single canonical correlation 0.3945.

5.12 Considered as a technique (like component analysis) of a purely mathematical kind for the convenient representation of data, canonical analysis has obvious advantages. Instead of operating on correlated data and disentangling the effects afterwards we attempt a disentanglement before the analysis begins by transforming to new variables with as much independence as possible. As in the component-analysis case, (and, indeed, as in ordinary scalar regression theory) we may then have to face the question whether our new variables have any obvious interpretation and can be identified with something "real", or whether they are to remain mere artefacts brought out by the mathematics. It would have been instructive at this point to give a number of practical examples of canonical analysis like those on factor and component analysis in Chapter 2. But, unfortunately, the technique has not yet been applied at all widely and there is a shortage of good illustrations. Theory, though far from complete, has outrun practice.

Example 5.3

(F. V. Waugh, 1942, *Econometrica*, **10**, 290.)

Waugh took, for each of the years 1921 to 1940 inclusive, the prices of beef steers and hogs and the per capita consumption of beef and pork (excluding lard) for the U.S.A. The prices were "deflated" by dividing by an index of per capita income, that is to say they purport to measure the changes in price relative to a stable value of money, and are given as dollars per 100 lbs. at Chicago. The consumption is given in pounds per annum.

We thus have, for 20 years ($n = 20$) a multivariate situation with $p = 2$, $q = 2$. We require to discuss the question how far meat consumption and meat prices are related, "meat" for this purpose including beef and pork but not mutton or chicken or minor sources of meat.

This, in one way, is a simplified form of the problem of canonical analysis because effectively we need only one linear

combination of the price – and consumption – variables and the greatest canonical correlation. The others are of minor interest. The correlation matrix was

	X_1	X_2	X_3	X_4
X_1 (steer prices)	1.	.18126	−.56396	−.49898
X_2 (hog prices)		1.	.35494	−.75671
X_3 (beef cons.)			1.	−.10293
X_4 (pork cons.)				1.

Let us note that these correlations make economic sense. The correlations between steer prices and beef consumption and between hog prices and pork consumption are negative; a rise in price means a fall in consumption. But the correlation between hog price and beef consumption is positive; when pork goes up in price there is a switch to beef, the consumption of which also goes up. (But the correlation between steer prices and pork consumption is negative so that substitutional effects are not entirely straightforward).

The classical way of discussing the question would probably have been to form an index-number (a weighted average) of prices and another (also a weighted average) of consumption and to investigate the relation between the two. The weights used in constructing these indices would have to be selected on prior grounds of a somewhat arbitrary kind: and the resulting correlation would, of course, depend on them. If we adopt the standpoint of canonical analysis the weights are determined for us by the condition that the correlation is a maximum. But we must always remember that there is nothing in the economics of the situation to compel the supposition that correlations are maximized. When we come to the stage of interpretation we shall encounter the same difficulty that we have already met in component analysis: of knowing whether our linear functions correspond to anything "real" or whether they are merely matters of mathematical convenience.

The greatest canonical correlation in this present example was −0.84666 (greatest, that is, in absolute value). The new variables, in terms of the standardized X's, were

$$\zeta_1 = \text{constant } (52.62\, X_1 + 47.38\, X_2)$$

$$\eta_1 = \text{constant } (25.38\, X_3 + 74.62\, X_4) \qquad (5.27)$$

where we have chosen the weights so as to add up to 100.

On looking at these values we see that the signs at least are acceptable. ζ_1 is an average of prices with nearly equal weights, a reasonable average for prices which both relate to the same quantity of meat; and the weights in η_1 are also both positive. Whether they are "reasonable" in the sense of providing a good index-number of the consumption of meat is another question, and one which has no answer unless we can say for what purposes the index is intended.

We may contrast this situation, which appears to give acceptable results, with the analysis of Example 5.2 where the canonical correlation was lower, 0.3945, and the canonical variables were

$$\zeta_1 = \text{constant } (-2.7722 X_1 + 2.2655\, X_2)$$

$$\eta_1 = \text{constant } (-2.4404 X_3 + X_4).$$

X_1 and X_2 were tests of reading (speed and power) and X_3 and X_4 were tests of arithmetic (speed and power). Since the coefficients in both the canonical variables are of opposite sign we cannot regard, say, ζ_1, as expressing some general capacity in reading formed by superposing speed and power. Had they been of the same sign we might well have suspected that the two were being combined into an index of ability to read. As it is, we seem led to the conclusion that speed and power are very different things and cannot be amalgamated in such a simple way as is embodied in the use of linear functions.

Example 5.4

(F. V. Waugh, *loc. cit.*)

The previous example should be regarded as illustrative only. The observations are correlated in time and it might be possible to obtain "nonsense" canonical correlations just as it is possible to obtain nonsense correlations in ordinary time-series analysis. The following example is free from this difficulty.

Information was available about 138 samples of Canadian Hard Red Spring wheat and the flour made from them. For the wheat five measurements were obtained:

X_1 kernel texture

X_2 test weight

X_3 damaged kernels

X_4 foreign material

X_5 crude protein in the wheat

For the flour there were four measurements

X_6 wheat per barrel of flour

X_7 ash in flour

X_8 crude protein in flour

X_9 gluten quality index

Here $n = 138$, $p = 4$, $q = 5$. The correlation matrix was

	1	2	3	4	5	6	7	8	9
1	1.	.75409	-.69048	-.44578	.69173	-.60463	-.47881	.77978	-.15205
2		1.	-.71235	-.51483	.41184	-.72236	-.41878	.54245	-.10236
3			1.	.32328	-.44393	.73742	.38132	-.54624	.17224
4				1.	-.33439	.52744	.46092	-.39286	-.01873
5					1.	-.38310	-.50494	.73666	-.14848
6						1.	.25056	-.48993	.24955
7							1.	-.43361	-.07851
8								1.	-.16276
9									1.

The first canonical correlation is ± 0.909,388, an unusually high value. The first two canonical variates are

$$\zeta_1 = \text{constant } (.35771\, X_1 + .29508\, X_2 - .56095\, X_3 - .44740\, X_4 + .50449\, X_5)$$
$$\eta_1 = \text{constant } (-X_6 - 0.53727\, X_7 + .84773\, X_8 + .04578\, X_9) \qquad (5.28)$$

where, as usual, the X's are expressed in standard measure. These are the (linear) index-numbers which give us the maximum correlation between the properties of the wheat and those of the flour.

Here again, having carried out the analysis, we need to look very carefully at the results to see if they make sense. On the whole it appears that they do. For example, in ζ_1 the signs given to kernel texture, test weight and crude protein are positive, these being the variables for which high scores indicate greater value; and those for the detrimental qualities of damaged kernels and foreign material are negative. In η_1 the wheat per barrel and ash content are negative and the crude protein and gluten quality positive. For these signs the canonical correlation is positive.

We could, of course, get equivalent results by changing all the signs of one of the canonical variables and taking the canonical correlation as −.909388. Which method of presentation we choose depends on the individual circumstances. The "constants" in (5.28) are positive.

It is, unfortunately, necessary to add that Waugh, to whom the analysis is due, carried out similar analyses on U.S. Hard Red Winter wheat, and found that although the canonical correlations were higher than for Canadian, the signs of the coefficients in the canonical variables were no longer satisfactory in all cases. Waugh suggests that the variables were dominated by X_5 and X_8, representing gluten content, which was inversely related to some of the other desirable characteristics.

It appears to me that in work of this character there may arise instabilities in the coefficients like those in Example 5.1. To decide the question we ought to work out the other canonical roots and see whether multicollinear effects are present. But the arithmetic would be formidable, although not beyond our capacity.

6. SOME PROBLEMS OF SAMPLING

6.1 In multivariate analysis, especially in those branches which deal with components, factors and canonical correlations, exact inferences from sample to parent offer some exceedingly difficult problems in distribution theory. In this chapter we shall attempt a sketch of what is known and what is unknown about these problems, but it will necessarily be a somewhat cursory survey. For a more detailed treatment of the mathematics reference may be made to Kendall's *Advanced Theory of Statistics*, volume 2, Wilks' *Mathematical Statistics* and Rao's *Advanced Statistical Methods in Biometric Research*. The difficulties are not, however, solely of a mathematical kind; some of them concern the proper model which should be set up in particular cases.

6.2 We begin with an account of the sampling distributions required. Practically everything that is known about exact distributions in the multivariate case depends on the assumption that the parent distribution is multivariate normal. This assumption will be made throughout.

It will be recalled that in samples from a univariate normal population the mean and variance are independently distributed. Also, in samples from a bivariate population the means are distributed independently of the variances and covariance. This type of result holds generally. In samples from a p-variate normal population the set of p means are distributed jointly normally, the correlation between $\bar{x}_i$ and $\bar{x}_j$ being ρ_{ij}, the correlation between x_i and x_j. The covariances and variances have a much more complicated form known as Wishart's distribution. Suppose we have a multivariate distribution with dispersion (covariance) matrix

$$(A_{ij})^{-1} = \rho_{ij} \qquad (6.1)$$

where we take the variates to have zero means and unit variances. (A simple transformation gives us the more general case if we need it.) The distribution is then

$$dF = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}p}} \exp\{-\frac{1}{2}\sum A_{ij}x_ix_j\}\prod dx \qquad (6.2)$$

Let the sample covariance be

$$a_{ij} = \frac{1}{n}\sum (x_i - \bar{x})(x_j - \bar{x}), \qquad (6.3)$$

where the summation is over the sample values. The joint distribution of the a's is then

$$dF = \frac{(\frac{1}{2}n)^{\frac{1}{2}p(n-1)} |A|^{\frac{1}{2}(n-1)} |a|^{\frac{1}{2}(n-p-2)}}{\pi^{\frac{1}{4}p(p-1)} \prod_{k=1}^{p} \Gamma\{\frac{1}{2}(n-k)\}} \exp\{-\frac{1}{2}n\sum A_{ij}a_{ij}\}\prod da \qquad (6.4)$$

Two points are to be noticed about this. First, the product $\prod da$ takes place over all a_{ij}, but not over both, for example, a_{12} and a_{21}. There are $\frac{1}{2}p(p+1)$ terms and more explicitly we should write the differential element

$$\prod_{j=1}^{p} \prod_{i \leqslant j}^{p} da_{ij}. \qquad (6.5)$$

On the other hand the summation in the exponential takes place over all i, j so that any term, say $A_{12}a_{12}$, occurs twice.

The second point is that n is the number in the sample and will always be used as such. Some writers take it as the "number of degrees of freedom" on the ground that the mean has been factorized off from the frequency function. This would involve putting n instead of $n-1$ in (6.4) and taking the sample number as $n+1$.

6.3 Theoretically we could get the distribution of sample correlation coefficients from (6.4) by integrating out the variances. This, however, leads to some intractable integrals (cf. the case when $p = 2$) and a good many distributional problems based on Wishart's distribution have to be tackled by roundabout methods, e.g. by finding lower moments and approx-

imating, or by reducing to the case where variates are independent.

In the case of independence we can obtain the distribution of sample correlations explicitly. The A_{ij} are then all zero for $i \neq j$ and the covariances do not appear in the exponential of (6.4). A simple variate transformation then gives us for the distribution of the r's

$$dF = \frac{|R|^{\frac{1}{2}(n-p-2)} \left[\Gamma \{\frac{1}{2}(n-1)\} \right]^{p-1}}{\pi^{\frac{1}{4}p(p-1)} \prod_{k=1}^{p} \Gamma\{\frac{1}{2}(n-k)\}} \prod_{i<j} dr_{ij} \qquad (6.6)$$

where $|R|$ is the sample correlation determinant $|r_{ij}|$. If $p = 2$ this reduces to the familiar form for the bivariate case.

6.4 The actual integration of (6.6) with respect to individual r's is complicated, not only because of the nature of R but because of the ranges of the r_{ij}. However, we can obtain from it the moments of the determinant $|R|$ by an argument due to Wilks. In fact

$$E\{|R|\}^t = \frac{\left[\Gamma \{\frac{1}{2}(n-1)\} \right]^{p-1}}{\left[\Gamma \{\frac{1}{2}(n-1) + t\} \right]^{p-1}} \frac{\prod_{k=1}^{p} \Gamma \{\frac{1}{2}(n-k) + t\}}{\prod_{k=1}^{p} \Gamma \{\frac{1}{2}(n-k)\}}. \qquad (6.7)$$

Bartlett, identifying the lower moments with those of a χ^2 distribution, obtained an approximation;

$$\chi^2 = -\{(n-1) - \frac{1}{6}(2p+5)\} \log |R|. \qquad (6.8)$$

is distributed as χ^2 with $\frac{1}{2}p(p-1)$ degrees of freedom.

6.5 An alternative and somewhat heuristic method of deriving the result, also due to Bartlett (1948, *Brit. J. Psych. Stat. Section*, 1, 73), is given by the consideration that

$$(1-r_{12}^2)\ (1-R_{3.12}^2)\ (\ .\ .\ .\ )\ (1-R_{p.12..(p-1)}^2)=|R| \qquad (6.9)$$

where the R's are multiple correlation coefficients. The various factors on the left here are independent in the case of parent independence. The logarithms of these factors multiplied by a factor in n are approximated by χ^2 distributions:

$$-\{n-\tfrac{1}{2}(3)\}\log(1-r_{12}^2) \qquad \text{with 1 d.f.}$$

$$-\{n-\tfrac{1}{2}(4)\}\log(1-R_{3.12}^2) \qquad \text{with 2 d.f.}$$

$$-\{n-\tfrac{1}{2}(p-1)\}\log(1-R_{p.12..(p-1)}^2) \qquad \text{with } p-1 \text{ d.f.}$$

Replacing the various factors in braces on the left by their weighted average $n-\frac{1}{6}(2p+5)$ we arrive at a χ^2 given by (6.8) with $\frac{1}{2}p(p-1)$ d.f.

6.6 It is generally true in this field that the distributions of determinants and determinantal ratios are easier to explore than those of their constituent items. They do not always give us the tests we should like, but in the present state of knowledge we have to be thankful for any test of even tangential relevance to the point under inquiry.

6.7 A distribution due to Hotelling bears the same relation to Wishart's as Student's t to the normal in the univariate case. Let $b_{ij}=n\,a_{ij}$ and let c_{ij} be the matrix inverse to b_{ij}. Then

$$T^2 = n(n-1)\sum c_{ij}\,\bar{x}_i\,\bar{x}_j \qquad (6.10)$$

is distributed in a form given by

$$dF = \frac{1}{B\{\frac{1}{2}(n-p),\ \frac{1}{2}p\}}\ \frac{\{T^2/(n-1)\}^{\frac{1}{2}(p-2)}}{(1+\frac{T^2}{n-1})^{\frac{1}{2}n}}\ d\left(\frac{T^2}{n-1}\right). \qquad (6.11)$$

5.8 One other set of distributional results is continually recurring in multivariate analysis. They concern the distribution of the characteristic roots of determinantal expressions in which the elements are covariances or analogous quantities. A great many branches of multivariate theory can be regarded as special cases of canonical regression or correlation and the basic results nearly all rest on a procedure which is common to them all. If L represents a vector of p quantities and A, B are $p \times p$ matrices, we require to maximize for variations in L the quadratic form $L'AL$ under the restriction that the other quadratic form $L'BL$ is constant. This, in fact, is effectively what we did in component analysis, with A as the correlation matrix and B as the unit matrix. Such problems lead to the solution of determinantal expressions of type

$$| A - \lambda B | = 0 \qquad (6.12)$$

In cases where one or both of A and B are subject to sampling fluctuations there will be generated a distribution of the p values of λ which are the roots of (6.12). Our interest centres in this distribution, and particularly in the distribution of largest, second largest ... roots.

6.9 Such distributions are exceedingly difficult to obtain. It is possible to find, in the normal case, the distribution of all roots together and of certain symmetric functions of them (in particular, their sum); but little exact knowledge is available about the individual roots except in the asymptotic case of large samples. This is a pity, because a thorough knowledge of the distributional situation would have very many applications. Apart from purely mathematical complexities the difficulties are three-fold.

(a) First, the number of parent parameters of a normal population rapidly increases with p, there being $\frac{1}{2}p(p+1)$ dispersion parameters apart from the p means. The number of possible nuisance parameters can therefore be extensive, and their removal by Studentization accordingly much more difficult to carry out.

(b) Secondly, the matrix B in (6.12) is sometimes

unknown and unobservable (or at least unobserved), as for example in the factor analysis case or the estimation of parameters in functional relations with several variables subject to error. Strictly speaking this is not a distributional problem; it arises from the model and the experimental situation; but it is none the less a severe handicap.

(c) Thirdly, in the particular case of characteristic roots we are faced with a problem which scarcely arises in ordinary distributional theory, namely the problem of identifying the variate we are discussing. If we imagine a set of values for A and B inserted in (6.12) and then allow A and B to vary slowly, we see that at certain points the roots in λ "change places", the one which was the second highest becoming, perhaps, the first highest. If the roots are all distinct and the sample is so large that the variation is small compared to the distance between them, they will retain their order and can be identified. This is one fundamental reason why we can make more progress in the asymptotic case, when the order of the roots is undisturbed by sampling effects although their values may alter.

6.11 An exact expression for the joint distribution of the roots is obtainable when the parent variates are all independent. The distribution can be put in various forms and we will give one of them purely as an illustration of the kind of form which arises.

In the distribution (6.4) suppose that all the parent covariances are zero and the parent variances unity and consider the roots of

$$|a - \lambda I| = 0 \qquad (6.13)$$

The determinant a is the product of the roots, say $\lambda_1 \ldots \lambda_p$, and the term in the exponential reduces to $-\frac{1}{2}n$ times the trace of a, which is the sum of the λ's. Consequently the frequency element of the joint distribution of λ's is

$$C\,(\lambda_1 \ldots \lambda_p)^{\frac{1}{2}(n-p-2)} \exp\Big(-\frac{1}{2}n \sum_{i=1}^{p} \lambda_i\Big), \qquad (6.14)$$

where C is a constant. The only difficulty arises from the evaluation of the differential element. This may be shown (e.g. Mood, 1951) to be

$$\prod_{i<j} (\lambda_i - \lambda_j) \prod d\lambda_i .$$

The evaluation of the constant can be carried out rather tediously. For our present purposes it is enough to note the form of the frequency element (the Fisher–Hsu–Roy distribution):

$$C\ (\lambda_1 \ldots \lambda_p)^{\frac{1}{2}(n-p-2)} \exp\left(-\tfrac{1}{2} n \sum_{i=1}^{p} \lambda_i\right) \prod_{i<j} (\lambda_i - \lambda_j),$$

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p \qquad (6.15)$$

6.12 Roy (1945 and many subsequent papers) has studied this distribution, the integral of which is a kind of generalized Beta-distribution. Rees and Foster (1957) have begun the tabulation of the percentage points of the *largest* root, the latter's work contemplating tables for $p = 2$, 3, 4 and 5.

An approximation due to P.L. Hsu (1941) states that the sum of the r smallest roots can be tested in the χ^2 distribution. More specifically if

$$\Lambda_r = n(\lambda_p + \lambda_{p-1} + \ldots \lambda_{p-r+1}) \qquad (6.16)$$

then Λ_r is distributed as χ^2 with $r(n-p+r)$ degrees of freedom. The λ's here are the roots of (6.13), not of the correlation matrix, i.e. of the dispersion matrix. Something of this kind is not unexpected for large samples. Each λ, in fact, is the variance of a principal component (when we standardize the covariance matrix to turn it into a correlation matrix) and the sum of r such, each with n d.f., would have $r\,n$ d.f. This remark is to be regarded as an observation on the consistency of the result, not as an unrigorous proof.

T.W. Anderson (1948, *J. Roy. Statist. Soc.* B, 10, 132) takes this a little further by showing that $(\Lambda_r - r\,n)/\sqrt{n}$ is asymptotically normal with mean zero and variance $2r$.

A very useful review is given in Anderson and Rubin (1956).

6.13 The quotation of results like this out of their context is, however, a little dangerous. The distributions and the tests based on them depend on the model under consideration and circumstances will alter the cases. This is only another way of saying, of course, that the test depends on the hypothesis; but the warning seems especially necessary in multivariate analysis.

Significance and estimation in component and factor analysis

6.14 In the component-analysis approach we do not postulate the existence of any "structure". We suppose that our observations x_{ij} are chosen at random from a p-variate population; that they are observed without error; and we narrow the sampling discussion by imposing the condition that this population is normal. Corresponding to any component analysis on a sample we may imagine an analysis carried out on the parent. Our sample values of λ are then estimators of parent values and our sample l's are estimates of parent values. It is meaningful to inquire how close the sample values lie to the parent values, to try to set confidence limits and so forth. This is the familiar situation of statistical inference where we are trying to estimate from sample to parent.

6.15 In such a case we are effectively looking for a variate transformation from x's to ζ's with the properties that the new variates are independent and that they account for as much variance as possible in descending order. Whether these variates can be identified with anything real or with a structure is not, at this stage, under examination. We may suppose, as we wish, that the individuals of the population are characterized by x's or by ζ's and the values which emerge for scrutiny in the sample are selected at random. In other words, we have to estimate the constants in

$$x_i = \sum_{k=1}^{p} l_{ik}\zeta_k \qquad (6.17)$$

but the ζ's are variates.

6.16 Except in the degenerate case when the variation lies in fewer than p dimensions, the parental characteristic equation of type $| R - \lambda I | = 0$ (where R now relates to parental coefficients and λ to parental values) is the limiting form of the sample equation. A simple appeal to continuity then suggests that the sample values are consistent estimators of the parent values.

Furthermore, the observed dispersion matrix in a normal distribution, say V, is easily seen to be a maximum-likelihood estimator of the parent dispersion matrix, say C; and individual variances of the sample are maximum-likelihood estimators of the parent values. We may then say that the sample correlation matrix r, as obtained from the sample dispersion matrix by division by the appropriate sample variances, is a maximum-likelihood estimator of the parent R; and in this sense the sample characteristic roots are the maximum-likelihood estimators of the parent values.

There is a pitfall to be avoided here. If we make a variate transformation

$$\zeta_i = \Sigma \beta_{ik} x_k \qquad (6.18)$$

to independent ζ's, the new frequency function gives a logarithm of the likelihood proportional to

$$-\tfrac{1}{2} n \Sigma \log \lambda_i - \tfrac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{p} \zeta^2_{ij} / \lambda_i, \qquad (6.19)$$

i.e. to

$$-\tfrac{1}{2} n \Sigma \log \lambda_i - \tfrac{1}{2} \sum_{j=1}^{n} \sum_{i,k,m} \beta_{ik} x_k \beta_{im} x_m / \lambda_i + | \beta_{ik} |^n. \qquad (6.20)$$

If we maximize this for variations in the λ's we get

$$\lambda_i = \sum_{k,m} \beta_{ik} \beta_{im} r_{km} \qquad (6.21)$$

but this is not a maximum-likelihood solution in the ordinary sense. The likelihood, in fact, is a constant under the rotation of (6.18). What we have done is to find the sample principal components.

6.17 This case is accompanied by one curious feature, namely that no non-zero λ requires any test of significance in this sense: a sample with d effective dimensions cannot arise from a population with fewer than d dimensions. However small a value of λ may be, therefore, it cannot arise from a parent value of zero. It may be negligible but it is always significant.

Nevertheless there is a sense in which we can "test the significance" of a *set* of λ's; or, better put, in which we can test their distinguishability. When we extract a number of components from a p-way complex there usually comes a time when the remaining λ's, though not vanishing, are relatively small. We may then ask ourselves: is the extraction worth continuing? This means that we suspect the remaining variation to be dimensionally isotropic – it could have arisen from a population in which all remaining λ's are equal, in which case there is no point in trying to maximize the variation in any particular direction. We therefore seek for a test of this property. Such an approximate test has been derived by Bartlett and its nature will be clear from one of his examples.

Example 6.1

(Bartlett, *Brit. J. Psych. Stat. Sect.*, 1950, 3, 77)

Hotelling has previously considered some data by Kelley obtained with four tests of reading speed, reading power, arithmetic speed and arithmetic power, carried out on 140 children. The correlation matrix was

$$\begin{bmatrix} 1 & .698 & .264 & .081 \\ & 1 & -.061 & .092 \\ & & 1 & .594 \\ & & & 1 \end{bmatrix}$$

The four roots of the characteristic equation were:

λ_1	1.846
λ_2	1.465
λ_3	0.521
λ_4	0.167
	3.999

and $|R| = 0.2353$.

A preliminary look at these results suggests that the first two components are meaningful; but there is a sharp drop from the second to the third and we are inclined to doubt the reality of the last two. Can we make this doubt into a precise hypothesis and test it?

In virtue of what we have already said, the third and fourth components must be "significant" if there is no error in the observations. Their importance is judged in the first instance from the magnitude of the observed λ's.

We may then argue in this way: we refer to the distribution of(6.6), that is to say the correlation distribution *in the case of parental independence*. The entire structure may be tested from (6.8) by testing $|R|$, or rather

$$-\{(n-1) - \tfrac{1}{6}(2p+5)\} \log |R| \qquad (6.22)$$

as χ^2 with $\frac{1}{2}p(p-1)$ d.f.

Here $|R| = 0.2353$, $p = 4$ and n is 140. Actually, for reasons which will appear presently, we take the factor of log $|R|$ in (6.6) as $n - p - \frac{1}{2} = 135.5$. We therefore have to test $-135.5 \log .2353 = 196.0$ with 6 d.f. This is highly "significant".

The interpretation is that such values could not have arisen from an uncorrelated parent. (Little is known of the power of such a test, but an intuitive judgment would suggest that its power is reasonably high against normally correlated

alternatives.) We therefore attribute "significance" to the first component.

Now arises the problem of extracting the first component and testing the residual; and so on. Bartlett proposes, after k roots $\lambda_1 \ldots \lambda_k$ have been accepted, to test the remainder by

$$\chi^2 = -\{n - 1 - \tfrac{1}{6}(2p + 5) - \tfrac{2}{3}k\} \log R_{p-k} \qquad (6.23)$$

with $\frac{1}{2}(p - k)(p - k - 1)$ d.f. where

$$R_{p-k} = \frac{|R|}{\lambda_1 .. \lambda_k \left[\frac{p - \lambda_1 \ldots . \lambda_k}{p - k}\right]^{p-k}} \qquad (6.24)$$

If n is large compared to p we can use the smallest, corresponding to $n - p - \frac{1}{2}$, for all k. This has the advantage that the values of χ^2 are additive. The point of (6.24) is that since $|R|$ is the product of all the λ's, R_{p-k} is the product of the remaining $p - k$, with a standardizing factor necessary to bring the sample variances back to unity. The assumption is that in "removing" the first k we are left with a $p - k$ complex of variation which we wish to test.

In the arithmetical example given,

$$R_2 = 0.521 \times 0.167 \times (2/0.689)^2 = 0.7330$$

$$R_3 = (0.2353 / 1.846) \times (3/2.154)^2 = 0.3444$$

$$R_4 = |R| = 0.2353.$$

For the tests of the λ's we then have

d.f.		
6	-135.5 (log .2353)	= 196.0
3	-135.5 (log .3444)	= 144.4
1	-135.5 (log .7330)	= 42.1

All the values are "significant" and we conclude that the λ's are effectively differentiated.

6.18 The test requires a modification when the analysis into components is carried out with standardization after evaluation of the latent roots (Bartlett, 1951, *Biometrika*, **38**, 337). In component analysis as we have expounded it we standardize the variates at the outset by dividing the x's by their respective sample variances. We could have proceeded without standardization by ascertaining the latent roots of $| V - \lambda I | = 0$ where V is the dispersion matrix; but this renders us dependent on the parent variances. Suppose we assume them equal but still work on the observed dispersion matrix. To standardize the values of λ we divide them by their mean, which is the same as the mean observed variance; the sum of the standardized λ's is then equal to p. In such a case

$$\chi^2 = -\{n - 1 - \tfrac{1}{6}(2p + 1 + \tfrac{2}{p})\} \log V \qquad (6.25)$$

is distributed approximately as χ^2 with $\frac{1}{2}(p + 2)(p - 1)$ degrees of freedom.

6.19 We may remark in passing certain results in the sampling theory of canonical correlations.

(a) Hotelling (1936) considered the large-sample theory of canonical roots. On the basis of an underlying multivariate population the correlation between two different roots is zero and the variance of a canonical correlation r_1 (the positive root of λ^2) can be tested as if it were an ordinary product-moment r by the formula

$$\text{var } r_1 = \frac{1}{n}(1 - \rho_1^2)^2 \qquad (6.26)$$

This suggests that Fisher's $\tanh^{-1}$ transformation would be useful in stabilizing the variance of canonical correlations and in normalizing the distribution, but the point does not seem to have been explored. We are assuming here, of course, that the roots of the characteristic determinant are all distinct.

(b) The distribution of the roots λ^2 of the determinant

(5.21) are of a form resembling that of the characteristic roots of an equation, with frequency element proportional to

$$\prod_{i=1}^{p} u_i^{\frac{1}{2}(q-p-1)} (1-u_i)^{\frac{1}{2}(n-q-p-2)} \prod_{i<j=1}^{p} (u_i - u_j) \prod_{i=1}^{p} du_i \qquad (6.27)$$

where $u_i = \lambda_i^2$ and it is assumed that there is no real correlation between the two groups of variates. In contradistinction to the principal components case this distribution gives a test of the reality (difference from zero) of the canonical correlations, not merely their separability.

(c) Corresponding to the regression of a scalar variate y on a set of x's we can regard canonical correlations or regressions as part of a theory generalizing the ordinary univariate theory. The natural generalization of the correlation between a p-way and a q-way vector is the vector correlation coefficient $K = \rho_1 \rho_2 \dots \rho_p$, the product of the canonical correlations; and the generalization of the alienation coefficient $1 - \rho^2$ of bivariate theory is the vector alienation coefficient $Z = (1-\rho_1^2)(\dots)(1-\rho_p^2)$. The sampling theory of both K and Z is easier than that of any one canonical correlation.

Estimation in factor-analysis models

6.20

Lawley's investigation

Lawley (Uppsala Symposium on Psychological Factor Analysis, March 1953) has discussed the large sample theory of estimation in factor-analysis, this work replacing some earlier investigations on the subject. He begins with the model (in my notation) as

$$x_i = \sum_{k=1}^{m} \alpha_{ik} \zeta_k + \varepsilon_i \qquad (6.28)$$

There are $m < p$ factors and the term ε here is a residual which may be either specific or an error of observation in

the x's. All we require of it is that it should be normally distributed with zero mean and variance σ_i^2. We assume also that the ratios of these variances are known (or, better still, their actual values). We may then re-scale the observations so that with new x's in (6.28) all the ε's may be taken to have the same variance σ^2. Two cases then arise, according as σ^2 is known or not known.

(The other case when the ratios of the error variances are unknown remains open for inquiry. I shall not discuss it here. There appears to be the same essential discontinuity between the known and unknown-variance-ratios situation as in the estimation of constants in functional relationships considered in chapter 4; and this is not surprising when we consider that we can eliminate the ζ's from p equations of type (6.28) to get a set of $p-m$ equations in the x's subject to errors ε.)

6.21 Lawley then considers the likelihood function and derives some results which we will briefly indicate without detailed proof.

(a) Let C be the population dispersion matrix and V the sample matrix. Let α_k denote the (column) vector $(\alpha_{1k}, \alpha_{2k}, \ldots \alpha_{pk})$ and circumflex accents denote maximum-likelihood estimators. Then in matrix notation

$$V\hat{\alpha}_k = \hat{\lambda}_k \hat{\alpha}_k \qquad (6.29)$$

$$\hat{\lambda}_k = \hat{\alpha}_k' \hat{\alpha}_k + \sigma^2 \qquad (6.30)$$

Thus we may take the m largest latent roots of V as estimators of λ. We assume that n is so large that the latent roots are distinguishable with negligible probability of being wrong.

(b) If, in addition, σ^2 also has to be estimated we have

$$(p-m)\hat{\sigma}^2 = \operatorname{tr} V - \sum_{k=1}^{m} \hat{\lambda}_k , \qquad (6.31)$$

or $\hat{\sigma}^2$ is the mean of the $p-m$ smallest roots.

(c) A criterion based on likelihood ratios may be derived to determine whether m factors ζ are enough. The criterion is

$$n \{ \log |\hat{c}|/|V| + tr(V\hat{c}^{-1}) - p \}, \qquad (6.32)$$

where $\hat{c}$ is C with the likelihood estimators inserted. This is distributed approximately as χ^2 with $\frac{1}{2}(p-m)(p-m-1)$ degrees of freedom.

But if σ^2 is also estimated the criterion is

$$n \log \{ |\hat{c}| / |V| \}, \qquad (6.33)$$

with $\frac{1}{2}(p-m-1)(p-m+2)$ degrees of freedom, one less than before.[2]

(d) The asymptotic variance of $\hat{\lambda}_k$ is

$$\text{var } \hat{\lambda}_k = \frac{2\lambda_k^2}{n} \qquad (6.34)$$

and asymptotically two different $\hat{\lambda}$'s are independent.

(e) The variance of $\hat{\sigma}^2$ is given asymptotically by

$$\text{var } \hat{\sigma}^2 = \frac{2}{n} \frac{\sigma^4}{p-m} \qquad (6.35)$$

(f) If σ^2 is known

$$\text{cov } (\alpha_{ik}, \alpha_{jk}) = \frac{\lambda_k}{n(\lambda_k - \sigma^2)} \left[c_{ij} - \frac{\lambda_k}{2(\lambda_k - \sigma^2)} \alpha_{ik} \alpha_{jk} \right.$$

$$\left. + \sum_{t \neq k} \frac{\lambda_t}{\lambda_k - \sigma^2} \left\{ \frac{(\lambda_k - \sigma^2)^2}{(\lambda_k - \lambda_t)^2} - 1 \right\} \alpha_{it} \alpha_{jt} \right] \qquad (6.36)$$

where c_{ij} is an element of C.

Also we have

$$\text{cov } (\alpha_{ik}, \alpha_{jt}) = \frac{-\lambda_k \lambda_t}{n(\lambda_k - \lambda_t)^2} \alpha_{it} \alpha_{jk}. \qquad (6.37)$$

(g) If σ^2 is not known we have to add to the right-hand side of (6.36) a term

$$\frac{\sigma^4}{2n(p-m)(\lambda_t-\sigma^2)^2}\,\alpha_{ik}\,\alpha_{jk} \qquad (6.38)$$

and to the right-hand side of (6.37) a term

$$\frac{\sigma^4}{2n(p-m)(\lambda_k-\sigma^2)(\lambda_t-\sigma^2)}\,\alpha_{ik}\,\alpha_{jt}\,. \qquad (6.39)$$

6.22 These results make it possible to apply at least rough tests in the factor-analysis case we are discussing. They are sufficiently complicated but not unmanageably so; and it is interesting to find expressions like $\lambda_k - \lambda_t$ appearing as denominators to mark the high sampling variability of results when two latent roots become close together. For small n, or for the case where the error variances have unknown ratios, much remains to be done.

6.23 We should also mention a different model from that of (6.28), discussed by Young (1941, *Psychometrika*, **6**, 49) and Whittle (1953, *Skand. Akt.* **35**, 223).

We write

$$x_i = \sum_{k=1}^{m} \alpha_{ik}\phi_k + \varepsilon_i, \qquad (6.40)$$

where the ε's are still stochastic error terms but α's and ϕ's are parameters under estimate. We now suppose there to be an underlying structure connecting the variables. Apart from the error term any x_{ij} is composed of a linear sum of products of two components. ϕ_{kj} is a quantity varying from individual to individual – we suppose that there are "factors" which for any variate i, appear in the individuals to different extents. The weights α are independent of the individual and measure the extent to which a factor appears in the ith variate. No hypothesis concerning the distribution of ϕ is required. As against this, we have more parameters to estimate. In Lawley's case (6.28) there are, apart from error

variance, p m parameters α. In the Young-Whittle case there are in addition m n ϕ's making $m(p + n)$ in all. There seems to be no obvious reason a priori why an analysis of this situation should bear any resemblance to the other.

6.24 It may be shown, however, that if we perform a least-squares analysis, i.e. minimize absolutely

$$\sum_j \sum_i (x_{ij} - \sum_k \alpha_{ik} \phi_{kj})^2 \qquad (6.41)$$

we do, on certain assumptions, arrive at the characteristic equation and the usual solution for the λ's. The validity of the least squares approach depends on an assumption that the error variances are all equal or that their ratios are known and hence that the measurements have been standardized so as to make the error variances equal. In this case the tests of "significance" of the latent roots and the error variances are different from those of Lawley's case.

I am bound to record my opinion, however, that the model represented by (6.40) is one which is not customarily required. And indeed, I do not understand its nature. Dr. Whittle and I have had a good deal of correspondence on the subject, at the end of which neither of us had succeeded in persuading the other to his point of view. The reader should be warned, therefore, that paragraphs 6.23 and 6.24 may inadequately present this part of the subject and should, if he is interested, pursue the topic in Dr. Whittle's papers.

6.25 This branch of the subject needs a good deal more investigation (a) to clear up confusions, (b) to obtain more effective moderate-sample tests and (c) to deal with tests after certain components have been extracted. Imperfect as may be the light we can throw on the subject, however, it is dazzling compared to the obscurity surrounding questions of significance in analysis by methods other than principal components such as centroid factor analyses.

6.26 The major question which has always exercised factor analysts (in these cases where the exact structure was not assumed beforehand) has been simply: when to stop factoring?

In psychological work there seems to have grown a divergence of practice between British and American psychologists. The former usually extract from two to four factors, the latter more and sometimes up to a dozen. P. E. Vernon, after an exhaustive discussion of sundry methods of deciding when to stop, came to the conclusion that they gave such discrepant results when applied to the same correlation matrices that it was doubtful whether any of them effectively covered the conditions of such analyses; and he went on to express a view that it may never be possible to specify the sampling errors of centroid and similar techniques precisely. It would be rash to suppose that no further progress is possible, if only by sampling experiments with modern computing equipment; but it certainly seems that the problem will not yield to a direct frontal approach of the classical statistical kind. Perhaps the most useful work which could be done in this field would be an investigation into the actual distribution function of (say) the four largest roots of the characteristic determinant on certain simple non-null hypotheses about the parent. Asymptotic results are hardly sufficient for the sample sizes which are customarily met with in practice.

7. NOTES ON THE HISTORY OF MULTIVARIATE ANALYSIS

7.1 The remaining chapters 8 and 9 of these notes will take us on to ground which, in a sense, is more familiar. It largely consists of generalizations to the multivariate case of the known results for univariate cases: analysis of dispersion, regression, tests of hypotheses and so forth. Not all these generalizations are straight-forward, and some of the multivariate theory, e.g. that relating to discriminant functions, is of a new type. But broadly speaking few new statistical ideas emerge. The main difference between uni- and multivariate analysis in such branches lies in the increasing complexity of the mathematics and the increasing difficulty of interpreting the results.

7.2 The purpose of this chapter is to give a brief historical account of the development of the subject over the past thirty years. This is one way of getting an insight into the intricacies of the work. I shall, for convenience, divide the notes under three headings (a) Wilks' criterion, (b) Discrimination and (c) Latent roots. The three topics are of course interrelated and sometimes an individual was writing on all three subjects at the same time; but the segmentation has reason as well as convenience to justify it.

7.3 Multivariate sampling-distribution theory may be regarded as beginning with the publication by Wishart in 1928 of the distribution known by his name. This gave for p normal variates the distribution of the variances and covariances which previously had been found by Fisher in 1917 for the bivariate distribution. It was latent in Wishart's work, and was brought out explicitly by Wilks in 1932, that the dispersion matrix (a_{ij}) was the natural extension to the multivariate case of the variance in univariate theory. Consequently one line

of development has been the study of *ratios* of matrices of the dispersion type, this being the multivariate analogue of the family of results (Student's t, analysis of variance, regression tests, etc.) which in univariate theory lead up to a test based on a variance ratio. A second line has been the study of differences of matrices of type $A - \lambda B$, or of maximization under constraint which, as we noted in chapter 6, lead to the study of determinantal roots. A third line has been the measurement of distance between p-variate populations and the associated study of discriminant functions. No one of these topics has been pushed as far as it can be or will be, but in all three theory is in some danger of outstripping practice; and one would expect that a point has been reached where many of the results already attained need further study in order that they may be reduced to the possibility of numerical application.

Note on the history of Wilks' criterion

7.4 Wishart's distribution has itself been generalized. T. W. Anderson and Girschik (1944) and Anderson alone (1946) have considered the 'non-central' distribution, that is to say the distribution of the sum of squares and cross products of a p-variate sample taken about some (fixed) point other than the sample means. The distributions are, naturally, rather complicated but there seems no reason why they should not be brought to numerical application if the labour were considered worth while.

7.5 One of the basic papers on determinantal ratios is that of Wilks (1932). Among other things Wilks found the distribution of the ratios of two sample dispersion determinants based on independent samples from identical populations (the analogue of the F-ratio); explicitly for $p = 2$ and in the form of an integral for greater values of p. By an ingenious argument of wide application he evaluated the moments in the general case. He then proceeded to study the distribution of the ratio of a dispersion determinant to one of its principal minors.

7.6 Approaching the subject from the viewpoint of the corre-

lation ratio Wilks then dealt with what would nowadays be regarded as an analysis of variance between and within classes in a one-way classification. Suppose we have k independent samples from the same p-variate population. Let (a_{ij}) be the dispersion matrix of all samples together and b_{ij} the function of means defined by

$$b_{ij} = \frac{1}{n} \sum_{a=1}^{k} n_a (\bar{x}_{ia} - \bar{x}_i)(\bar{x}_{ja} - \bar{x}_j) \qquad (7.1)$$

where n_a is the number in the ath sample, $n = \Sigma n_a$, $\bar{x}_{ia}$ is the mean of the ith variate in the ath sample and $\bar{x}_i$ is the mean of the ith variate in all samples together. Let

$$c_{ij} = a_{ij} - b_{ij} \qquad (7.2)$$

so that (c) is a "within-class" dispersion matrix. Then the ratio

$$W = \frac{|c_{ij}|}{|a_{ij}|} \qquad (7.3)$$

$$= \frac{|c_{ij}|}{|b_{ij} + c_{ij}|} \qquad (7.3a)$$

is the natural extension of the ratio of "within-class" to "total" variance. The ratio arises more naturally in this form than as a "between-class" to "within-class" ratio $|b_{ij}| / |c_{ij}|$.

7.7 Wilks derived the explicit distribution of the ratio W (now generally known as Wilks' criterion) for $p = 1$ and $p = 2$, and the moments of the distribution for general p, in the case where the criterion arises as a likelihood ratio for testing the homogeneity of a set of means in samples from populations with identical dispersions. Finally he went on to consider the ratios of certain correlation determinants (as distinct from covariance determinants) and derived their moments. Later in 1935, he used similar ideas and methods to test the independence of k sets of normal variates.

7.8 Lawley (1938) also discussed the generalization of the F-ratio and proposed a test-function of a different kind from Wilks', namely

$$V = \sum_{i,j} c^{ij} b_{ij} \qquad (7.4)$$

where (c^{ij}) is the inverse of c_{ij}. Shortly afterwards the theory of characteristic or latent roots began its development and almost at once Hsu (1940) linked it up with the Wilks and Lawley criteria. In fact, if the non-zero roots of

$$| b_{ij} - \lambda (b_{ij} + c_{ij}) | = 0 \qquad (7.5)$$

are $\lambda_1, \ldots \lambda_m$, then

$$W = \prod_{i=1}^{m} (1 - \lambda_i) \qquad (7.6)$$

$$V = \sum_{i=1}^{m} \frac{\lambda_i}{1 - \lambda_i} \qquad (7.7)$$

It is the dependence of the Wilks-Lawley criteria on symmetric functions of the λ's and not on individual λ's which makes their sampling properties relatively easy to investigate. In a further paper (1941d) Hsu showed how the general regression problem could be reduced to a canonical form and that significance tests depend essentially on the distribution of W. Since Hotelling's T also depends on W this was equivalent to demonstrating an extension of Student's t to the testing of multivariate regression coefficients.

7.9 In the meantime Bartlett (1938) gave the first of a useful series of approximations to the distribution of W in the null case by fitting a χ^2 distribution to the lower moments. Hsu (1940) showed that nV and $-n \log W$ tend to be certainly identical for large n and that both have a distribution which can be expressed as a non-central χ^2. Wald and Brookner (1941) also gave an expression for the distribution of $-\log W$ and made a beginning with the tabulation of percentage points in special cases.

7.10 Progress was being made rapidly at this time. It was

checked by the war but not for long. In 1946 Wilks extended his earlier work to the testing of equality of means, variances or covariances in a multivariate normal system. He and Tukey (1946) approximated to the distribution by fitting a Beta distribution (not, as in Bartlett's case, a Gamma-distribution) to the lower moments and Wilks tabulated some of the percentage points. In the same year (1946) T. W. Anderson examined the distribution of the ratio W when the dispersion elements are distributed in a non-central Wishart form and obtained the moments for $p = 1$ and 2. (There seems on many occasions to be a natural barrier at $p = 2$ preventing extensions to higher p without the importation of more complicated transcendental functions.) Rao (1948b) later examined Bartlett's approximation.

7.11 The problem of nuisance parameters has also received some attention. Plackett (1947), following some ingenious work by Pitman and Morgan in devising tests of equality of variances in bivariate populations independent of parental covariances, derived an exact test for the equality of variances or covariances in any number of uni- or bi-variate populations and gave extensions to pairs of three- or four- variate populations subject to certain conditions on the sample size. Recently G. S. James (1954) has considered the extension of Welch's work in testing the difference of means where variances are unknown.

A historical note on latent roots

7.12 We have noted how criteria based on determinantal ratios depend on symmetric functions of latent roots of certain matrices. Occasions often arise in which we want the distribution of one of the roots, which raises some new and complex distributional problems. The necessity for this knowledge arises in two main ways:

(a) in the reduction of multivariate situations to a 'standard' or 'canonical' form, as for instance in component and canonical correlation analysis;

(b) in testing for the equality of two matrices, as for instance by examining whether a root of $|A - \lambda B| = 0$ is

near unity. This is the approach which has been developed by S. N. Roy since 1939.

7.13 Let us note the formal equivalence of certain customary ways in which the problem arises. If (b) and (c) are matrices of the dispersion type independently distributed in the Wishart form the canonical correlations are, in effect, the roots of

$$| \lambda (b_{ij} + c_{ij}) - b_{ij} | = 0. \qquad (7.8)$$

We can, of course, write

$$\mu = \frac{\lambda - 1}{\lambda}$$

and derive (apart from zero roots)

$$| c_{ij} + \mu b_{ij} | = 0. \qquad (7.9)$$

If we change the sign of μ and regard the sample number associated with b_{ij} as tending to infinity; and if we let the variances of b be unity and the covariances zero we get

$$| c_{ij} - \lambda I | = 0. \qquad (7.10)$$

And, subject to the element of approximation induced by standardizing *sample* values of c_{ij} so as to have unit variances we then derive

$$| r_{ij} - \lambda I | = 0. \qquad (7.11)$$

7.14 The large-sample theory of canonical correlations was worked out by Hotelling (1936). There does not seem to have been much more done in this field until the recent work by Lawley (referred to in chapter 7) on factor analysis. It would be useful to have this topic explored further.

7.15 Attention has been concentrated more on exact distributions in the case of normal variates and, as might be expected the greatest progress has been made in the null case, i.e. where the p variates are independent. The basic distribution here

is the Fisher–Hsu–Roy distribution given for a special case in chapter 6. This gives the simultaneous distribution of the roots but for p greater than 2 it is a difficult distribution to handle because of the complicated range of variation of the successive λ's, which render almost impossible the integrating out of unwanted roots. The greatest progress here has been made by Roy (1945) and Nanda (1948) in dealing with the case of (7.9) where the (a) and (b) matrices emanate from identical populations.

7.16 For the non-null case expressions have been derived by Bartlett (1947) giving the distribution of canonical correlations in the form of an infinite series rather like Fisher's series for multiple R in the non-null case. (Hsu, 1941a, had obtained the limiting form of the distribution). The distribution is tractable when there is only one non-vanishing parent canonical correlation, but is stubborn for more than one non-vanishing parental correlation.

7.17 Little attention seems to have been given to the distribution of the vectors corresponding to the latent roots but reference may be made to T. W. Anderson (1951 b) on this subject.

A historical note on discriminatory analysis

7.18 Doubtless the idea of discriminating between multivariate populations could be traced far back into the past. For present purposes the history of the subject may be regarded as beginning with the work of Karl Pearson round about 1920. Pearson, considering anthropometric data, was led to seek a coefficient which would, in some acceptable sense, "measure the distance" between two populations. The first work to be published on his coefficient of racial likeness, which he denoted by C^2, was Miss Tildesly's on Burmese skulls (1921, *Biometrika*, **13**, 247).

7.19 About the same time P. C. Mahalanobis developed an interest in the subject and came to the conclusion that Pearson had not achieved his object. C^2 varied very much with the sample number and, although it provided a test of significance,

it did not measure the magnitude of the difference between two populations. Mahalanobis accordingly proposed an alternative measure which he called D^2 and used it in 1925 to discuss racial mixtures in Bengal. This far-sighted work was the starting point of research by the Indian school which is still in progress.

7.20 If two multivariate normal populations have the same dispersion matrix (α_{ij}) with an inverse (α^{ij}) and means μ_{i1}, μ_{i2} $(i = 1, 2, \ldots p)$ the distance between means of variates may be defined as $\delta_i = \mu_{i1} - \mu_{i2}$ and Mahalanobis' generalized distance for the population may be written

$$\Delta^2 = \sum_{i,j} \alpha^{ij} \delta_i \delta_j. \qquad (7.12)$$

A corresponding formula holds for the sample values:

$$D^2 = \Sigma a^{ij} d_i d_j. \qquad (7.13)$$

where

$$d_i = \bar{x}_{i1} - \bar{x}_{i2}.$$

If the variates are independent (7.12) reduces to

$$\Delta^2 = \Sigma \frac{\delta_i^2}{\text{var } x_i} \qquad (7.14)$$

and if we scale the variates so as to have unit variances this becomes the square of the 'distance' between the parent means in the customary sense. We may also regard (7.12) as defining an ordinary distance in a Euclidean space with oblique axes. It follows that, given three populations P, Q and R, the distance between P and Q is not greater than the sum of distances P to R and R to Q.

7.21 It is to be emphasized that the definition applies only to normal populations with identical dispersion matrices. In the language of differential geometry, it is based on a

Euclidean metric, not a Riemannian metric. In other cases it may be unsuitable.

7.22 Between 1927 and 1930 Mahalanobis had some controversy with K. Pearson, who defended C^2. Mahalanobis (1930, *Jour. Asiatic Soc. Bengal*, 26, 541) continued his practical and theoretical research on D^2. Pearson, though writing on his coefficient as late as 1936 just before his death, failed to find much support for his views.

7.23 At this point Hotelling (1931) generalized "Student's" t. His T was, in fact, equivalent to Mahalanobis' D^2, for

$$T^2 = \frac{D^2 \, n_1 \, n_2}{(n_1 + n_2)\,(n_1 + n_2 - 2)} \qquad (7.15)$$

where n_1 and n_2 are the two sample numbers. It was some time, however, before this equivalence was realized, or before it was pointed out that in the null case the distribution of both was equivalent to the distribution of the multiple correlation coefficient R^2 in the null case. In the meantime Fisher (1930) obtained the distribution of R^2 in the non-null case.

7.24 There appears to have been a lull in this particular field of statistical theory for five or six years during which the theory of testing hypotheses, beginning in 1928, was developed by by Neyman and E. S. Pearson. In 1936 Fisher published his first paper on discriminant functions and resolved the problem of testing significance. The main difference between his approach and that of Mahalanobis was that the latter was measuring distance whereas Fisher was merely concerned to divide the sample space into two regions and allocate a sample value to one population or another according to which region it fell into; to that extent his approach was nearer to the modern theory of decision functions. But again Mahalanobis' D^2 appeared in a natural role and the approaches were obviously very close.

7.25 The distribution of D^2 in the non-null case when both populations have a known dispersion matrix was found by R. C. Bose in 1936. Two years later he and Roy "Studentized" the distribution. Fisher (1938) about the same time gave what is

effectively a generalization of D^2 for more than two populations. And Welch (1939) linked up the theory of discriminatory functions and that of statistical tests in a simple but effective way.

7.26 There is here a break in continuity which is not entirely due to war. The English school on the whole moved over to the study of latent roots and canonical correlations or concerned themselves with the practical problems of applying discriminant functions (Penrose, 1947 and C. A. B. Smith, 1947). Similar studies began in the U.S.A., as for example with the work of Cochran (1943), von Mises (1945), Cochran and Bliss (1948) and T. W. Anderson (1951 a). Further theoretical developments are due to Roy and Rao – see in particular Rao (1946–1948 a, 1948 c, 1949, 1950).

7.27 Rao (1948 c) made some interesting comments on the general problem of measuring distance which do not appear to have been followed up, possibly because tensor analysis and differential geometry are not very familiar fields to most mathematical statisticians.

Suppose we have a multivariate population depending on u parameters $\theta_1 \ldots \theta_u$ with density function $\phi(x, \theta_1, \ldots \theta_u)$, where for short we write x to denote all the variables. We may set up a parameter space of u dimensions and consider two neighbouring points in it typified by θ and $\theta + d\theta$. The differences between the probability densities of the two points may be written $d\phi$, where d relates to variations in θ. Let us now agree to measure the discrepancy between the populations represented by the two points by their relative difference $d\phi / \phi$. The variance of this quantity seems a reasonable measure of discrepancy and becomes

$$ds^2 = \sum_{i,j} g_{ij}\, d\theta_i\, d\theta_j, \qquad (7.16)$$

where

$$g_{ij} = E\left(\frac{1}{\phi}\frac{\partial \phi}{\partial \theta_i}\right)\left(\frac{1}{\phi}\frac{\partial \phi}{\partial \theta_j}\right), \qquad (7.17)$$

namely, what is usually called the information matrix. The

form ds^2 is invariant and the g's are consequently a covariant tensor of the second order. We may go further and regard ds^2 as the quadratic differential matrix defining the element of length.

7.28 Given, then, two points A and B (not necessarily close together) in the parameter space we can find the distance between them by integrating ds^2 along a geodesic. If the equations of this geodesic are

$$\theta_i = f_i(t), i = 1,2, \ldots u \qquad (7.18)$$

where t is a variable, the functions θ are solutions of the equations

$$\sum_{j=1}^{u} g_{ik} \frac{d^2\theta}{dt^2} + \sum_{j=1}^{u} (j\ l,\ k) \frac{d\theta_j}{dt} \frac{d\theta_l}{dt} = 0 \qquad (7.19)$$

where $(j\ l,\ k)$ is the Christoffel symbol defined by

$$(j\ l,\ k) = \tfrac{1}{2} \left(\frac{\partial g_{jk}}{\partial \theta_l} + \frac{\partial g_{lk}}{\partial \theta_j} - \frac{\partial g_{jl}}{\partial \theta_k} \right). \qquad (7.20)$$

Theoretically the g's are determinable from (7.17) and hence the θ's by the solution of these equations. We can then find the distance by integrating ds^2 from A to B along the curve defined by (7.18).

This would be an attractive measure of distance, being invariant under very general transformations of the parameters, but the *estimation* of the distance raises some difficulties. Mahalanobis' distance is a particular case, for the geodesics then become straight lines.

7.29 Rao, following Bhattacharyya (1946) also suggested representing populations on a unit hypersphere, the distance between them then being the angle

$$\arccos \int \{f(x)\ \psi\ (x)\}^{\frac{1}{2}}\ dx \qquad (7.21)$$

where f, ψ, are the frequency functions of the two populations.

This approach and that of the previous paragraph are both free from any restriction as to the shape of the frequency function or the number of parameters involved.

8. TESTS OF HOMOGENEITY

8.1 Throughout this chapter we shall suppose that the variation is normal. We shall consider samples from k different p-variate populations and three types of hypothesis:

H : That the populations have the same means, the same variances and the same covariances;

H_1 : That the populations have the same variances and the same covariances, irrespective of the means;

H_2 : That, *given* the equality of variances and covariances, the means are equal.

The results are a natural extension of homogeneity tests for k univariate normal samples based on likelihood ratios (for which see Kendall, *Advanced Theory of Statistics*, vol 2). The hypothesis which generalizes the ordinary analysis of variance is H_2 which, as a general rule, leads to criteria with simpler distribution functions than those of H and H_1.

The Pearson-Wilks results for bivariate populations

8.2 We consider first the case of k bivariate populations ($p = 2$). The tth population has means μ_t and ν_t for variates x and y, and the dispersions are σ^2_{xt}, $\sigma_{xt}\,\sigma_{yt}\,\rho_t$ and σ^2_{yt}. We draw a sample of n_t members from it, the total

$$\sum_{t=1}^{k} n_t$$

being n. Consider the hypothesis H. We ascertain two likelihoods, $P(\Omega)$, the maximum likelihood when all the parameters

are different, and $P(\omega)$, the maximum likelihood under the restrictions of H that means and dispersions are common to the populations.

If the likelihood for the first case is written down and differentiated with respect to the various parameters we find

$$\hat{\mu}_t = \bar{x}_t , \quad \hat{\nu}_t = -y_t, \tag{8.1}$$

$$\hat{\sigma}_{xt} = s_{xt} , \quad \hat{\sigma}_{yt} = s_{yt} , \quad \hat{\rho}_t = r_t, \tag{8.2}$$

where $\bar{x}_t$ and $\bar{y}_t$ are sample means and s_{xt}, s_{yt}, r_t the sample standard deviations and correlation. (In arriving at the standard deviations we use divisors of types n_t, not $n_t - 1$.)

Similarly, for the second case we find

$$\hat{\mu} = \bar{x}_0, \quad \hat{\nu} = \bar{y}_0, \tag{8.3}$$

$$\hat{\sigma}_x^2 = \nu_{110} = \nu_{11a} + \nu_{11m}, \tag{8.4}$$

$$\hat{\sigma}_y^2 = \nu_{220} = \nu_{22a} + \nu_{22m}, \tag{8.5}$$

$$\hat{\sigma}_x \hat{\sigma}_y \hat{\rho} = \nu_{120} = \nu_{12a} + \nu_{12m}, \tag{8.6}$$

where

$$\bar{x}_0 = \frac{1}{n} \sum_{t=1}^{k} n_t \bar{x}_t \tag{8.7}$$

$$\bar{y}_0 = \frac{1}{n} \sum_{t=1}^{k} n_t \bar{y}_t \tag{8.8}$$

$$n \nu_{110} = \sum_{t=1}^{k} \sum_{u=1}^{n_t} (x_{tu} - \bar{x}_0)^2 = n s_{x0}^2 , \text{ say}, \tag{8.9}$$

$$n \nu_{120} = \Sigma \Sigma (x_{tu} - \bar{x}_0)(y_{tu} - \bar{y}_0) = n s_{x0} s_{y0} r_0, \text{ say}, \tag{8.10}$$

$$n \nu_{220} = \Sigma \Sigma (y_{tu} - \bar{y}_0)^2 = n s_{y0}^2 , \text{ say}. \tag{8.11}$$

Thus s_{xo}^2 , s_{yo}^2 and r_o are the variances and correlation of the pooled samples. Further

$$n_t \nu_{11a} = \sum_{t=1}^{k} \sum_{u=1}^{n_t} (x_{tu} - \bar{x}_t)^2 = \sum_{t=1}^{k} n_t s_{xt}^2 \qquad (8.12)$$

$$n_t \nu_{11m} = \sum_{t=1}^{k} n_t (\bar{x}_t - x_o)^2 \qquad (8.13)$$

with corresponding expressions for ν_{12a}, ν_{22a} etc. We also write for each sample

$$n_t \nu_{11t} = \sum_{u=1}^{n_t} (x_{tu} - \bar{x}_t)^2 = n_t s_{xt}^2, \qquad (8.14)$$

$$n_t \nu_{22t} = \sum_{u=1}^{n_t} (y_{tu} - \bar{y}_t)^2 = n_t s_{yt}^2 \qquad (8.15)$$

$$n_t \nu_{12t} = \sum_{u=1}^{n_t} (x_{tu} - \bar{x}_t)(y_{tu} - \bar{y}_t) = n_t s_{xt} s_{yt} r_t \qquad (8.16)$$

If we substitute the appropriate values in the original likelihoods we find

$$\lambda_H = \frac{P(\omega)}{\rho(\Omega)} = \prod_{t=1}^{k} \left\{ \frac{|\nu_{ijt}|}{|\nu_{ijo}|} \right\}^{\frac{1}{2} n_t} \qquad (8.17)$$

where

$$|\nu_{ijt}| = \begin{vmatrix} \nu_{11t} & \nu_{12t} \\ \nu_{12t} & \nu_{22t} \end{vmatrix} = s_{xt}^2 s_{yt}^2 (1 - r_t^2) \qquad (8.18)$$

$$|\nu_{ijo}| = \begin{vmatrix} \nu_{110} & \nu_{120} \\ \nu_{120} & \nu_{220} \end{vmatrix} = s_{xo}^2 s_{yo}^2 (1 - r_o^2). \qquad (8.19)$$

These are, in fact, the generalized variances of the tth sample and of all samples together, and the likelihood ratio thus appears as a product of Wilks criteria.

8.3 In an analogous way we find

$$\lambda_{H_1} = \prod_{t=1}^{k} \left\{ \frac{|\nu_{ijt}|}{|\nu_{ija}|} \right\}^{\frac{1}{2} n_t} \qquad (8.20)$$

$$\lambda_{H_2} = \left\{ \frac{|\nu_{ija}|}{|\nu_{ijo}|} \right\}^{\frac{1}{2} n} \qquad (8.21)$$

and we note that, as in the univariate case

$$\lambda_H = \lambda_{H_1} \times \lambda_{H_2}. \qquad (8.22)$$

We may conveniently indicate by suffices 1 and 2 the criteria appropriate to H_1 and H_2.

8.4 Our next problem is to find the sampling distributions of these criteria. This is performed by Wilks' method (we shall omit the details) of finding the moments of the λ's, or rather, of $L = \lambda^{2/n}$, which will serve equally well as a criterion because it is monotonically dependent on λ. The distribution of L_2 is fairly simple, being given by

$$dF = \frac{\Gamma(n-2)}{\Gamma(n-k-1)\,\Gamma(k-1)} (\sqrt{L_2})^{n-k-2} (1-\sqrt{L_2})^{k-2} \, d\sqrt{L_2} \qquad (8.23)$$

The distributions of L_1 and L (relating to H) are more complicated except when $k = 2$ and approximations have to be employed.

8.5 We may also note that the L-criteria have certain properties to recommend them on intuitive grounds. They vary from 0 to 1 and as they decrease from unity we are more inclined to reject the corresponding hypothesis, i.e. small values are significant. Consider, for example

$$(\lambda_{H_2})^{2/n} = \frac{|\, v_{tja} \,|}{|\, v_{tja} + v_{tjm} \,|} \qquad (8.24)$$

For this to be unity (v_{tjm}) must be zero and all the sample means are the same. As they move further apart λ_{H_2} decreases; it is zero if and only if one of the differences of means is infinite or if within-class variances are zero or $r_t = 1$ for any t. Furthermore it decreases monotonically, that is to say, has no other maximum apart from unity.

Example 8.1

(Pearson and Wilks, 1933)

Five samples are available, each of twelve members, of aluminium die-castings. ($k = 5$, $n_t = 12$ for all t, $n = 60$). On each specimen two measurements are taken; tensile strength (1000 lb. per square inch) which we call x, and hardness (Rockwell's E) which we call y. The data may be summarized as follows

Sample Number t	x		y		Correlation Coefficient
	Mean	Standard Deviation	Mean	Standard Deviation	
1	33.399	2.565	68.49	10.19	0.683
2	28.216	4.318	68.02	14.49	0.876
3	30.313	2.188	66.57	10.17	0.714
4	33.150	3.954	76.12	11.18	0.715
5	34.269	2.715	69.92	9.88	0.805

We are interested in the homogeneity of these data.

We first of all test H_1, that the data show no significant difference in dispersions. We have the following results

t	Sums of Squares nv_{11t}	Sums of Squares nv_{22t}	Sums of Products nv_{12t}	Generalized Variances $\lvert v_{ijt} \rvert$	$\log_{10}\lvert v_{ijt} \rvert$
1	78.948	1247.18	214.18	365.204	2.56254
2	223.695	2519.31	657.62	910.401	2.95923
3	57.448	1241.78	190.63	243.029	2.38566
4	187.618	1473.44	375.91	938.451	2.97241
5	88.456	1171.73	259.18	253.281	2.40360
Totals	636.165	7653.44	1697.52		13.28344

Hence we find

$$v_{11a} = \frac{1}{60}(636.165)$$

$$= 10.6028$$

$$v_{22a} = 127.5573$$

$$v_{12a} = 28.2920$$

$$\lvert v_{ija} \rvert = 552.018$$

$$\log L_1 = \{1/k \log \Pi \lvert v_{ijt} \rvert - \log \lvert v_{ija} \rvert\}$$

$$= \bar{1}.914,734$$

giving $L_1 = .8217$

We test, generally, H with $5(k-1)$ degrees of freedom, H_1 with $3(k-1)$ and H_2 with $2(k-1)$. There is a useful

general theorem (due to Wilks) of which a particular case is that for large samples $-n \log_e L_1$ is distributed as χ^2 with a number of degrees of freedom equal to the number of constants fitted in Ω less the number fitted in w, in this case $3(5-1) = 12$. This gives 11.78 as χ^2 with 12 d.f. which is not significant. A more exact test confirms this.

We therefore accept the equality of dispersion and proceed to test the means by hypothesis H_2. We now have a generalized analysis of variance.

	D.F.	S.S. (x)	S.S. (y)	S.P. (x,y)	Generalized Variances
Between samples	$k-1=4$	306.089	662.77	214.86	$\lvert\nu_{ijm}\rvert=43.528$
Within samples	$n-k=55$	636.165	7653.42	1697.52	$\lvert\nu_{ija}\rvert=552.018$
Totals	$n-1=59$	$n\nu_{110}=942.254$	$n\nu_{220}=8316.19$	$n\nu_{120}=1912.38$	$\lvert\nu_{ijo}\rvert=1160.77$

$$L_2 = \{ \lvert\nu_{ija}\rvert / \lvert\nu_{ijo}\rvert \} = 0.4750$$

We can apply an approximate test to $-n \log L_2$, in the same way. The quantity is 44.59 with 13-5=8 d.f. This is highly significant. In this particular case an exact test is available, the ratio

$$\frac{1-\sqrt{L_2}}{\sqrt{L_2}} \qquad \frac{n-k-1}{k-1}$$

being distributed as an F ratio with $2(k-1)$ and $2(n-k-1)$ degrees of freedom. This also rejects the hypothesis, the ratio being 4.95 with 8 and 108 d.f. (We shall discuss later the cases in which exact F-ratio tests are available.)

We conclude that there is heterogeneity in the mean values. We therefore proceed to test x and y separately:

	Estimates of variance		d.f.
	x	y	
Between samples	76.522	165.69	4
Within samples	11.566	139.15	55

A simple F ratio test shows that at the 1% point the differences between mean strengths are significant, but not those between hardness.

The conclusion is that although the samples are under control as regards dispersions and hardness they are not under control as regards mean-strength, although hardness and strength are fairly highly correlated. One would, I think, be led to consider the accuracy and nature of the methods of measurement before drawing conclusions about the materials themselves.

Example 8.2

(Pearson and Wilks, 1933)

Measurements were available on the lengths and breadths of 600 skulls, 20 from each of 30 races. (The details are given in the paper under reference). That there would be some variation between races was to be expected but it was of interest to see whether it extended to dispersions. The hypothesis H_1 was tested. The authors found

$$|\nu_{ija}| = 656.369, \quad 1/k \sum \log \{|\nu_{ijt}|\} = 2.644429$$

$$\log |\nu_{ija}| = \underline{2.817148}$$

$$\text{Difference} = \bar{1}.827281$$

$$L_1 = 0.6719$$

The simple test gives 600 (.39765) = 236.8 distributed as χ^2 with 150-63 = 87 d.f. This is highly significant.

There is therefore lack of uniformity in the dispersions. We now proceed to consider them individually.

The homogeneity of a set of variances in the univariate case can be tested by known methods, which are, in fact, simpler forms of the likelihood criteria we are now employing

here. It is found that the variances of both x and y are significantly heterogeneous. Finally we consider the correlations r. A convenient test of homogeneity here is obtained by transforming the 30 observed r's by the formula $z = \tanh^{-1} r$ and testing

$$\chi^2 = \sum_{t=1}^{30} (n_t - 3)(z_t - \bar{z})^2 \qquad (8.25)$$

with 29 degrees of freedom. In the present instance this led to $\chi^2 = 96.01$ which is very significant.

The general conclusion is that there exists racial heterogeneity in variances and correlation, the latter being apparently less uniform than the variances themselves.

In a case of this kind we should probably not wish to proceed to test heterogeneity of means; but if we did so wish, we should proceed as follows: since H_1 is untenable we should not test H_2. We revert to tests of type H_1 on the variates x and y separately. We can then proceed on the basis of univariate theory, by the Behrens-Fisher test, or by assuming that heterogeneity is not important enough to invalidate an ordinary F-ratio test.

8.6 Methods of a parallel kind could be followed for the testing of k samples of p-variate populations, although I am not aware that the general case has been worked out explicitly. Wilks (1946) has, however, considered a problem of a very similar character for the p-variate case. We are given a sample of n from *one* p-variate population with means μ_i and dispersions $\rho_{ij}\sigma_i\sigma_j$. We consider the hypotheses

H: that the means are all equal and the corresponding dispersions all equal for each variate;

H_1: that the dispersions are equal regardless of the means;

H_2: that, given the equality of dispersions, the means are all equal.

Let

$$\bar{x}_i = 1/n \sum_{j=1}^{n} x_{ij} \tag{8.26}$$

$$\bar{x} = 1/p \sum_{i=1}^{p} \bar{x}_i \tag{8.27}$$

$$s_{ij} = 1/n \sum_{t=1}^{n} (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j) \tag{8.28}$$

$$s^2 = 1/p \sum_{i=1}^{p} s_{ii} \tag{8.29}$$

$$s^2 r = \frac{1}{p(p-1)} \sum_{i \neq j=1}^{p} s_{ij}. \tag{8.30}$$

The hypothesis H to be tested is

$$\mu_i = \mu, \tag{8.31}$$

$$(\rho_{ij}\sigma_i\sigma_j) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \cdot & \cdot & \cdots & \cdot \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \tag{8.32}$$

and the likelihood is

$$P = \frac{|\alpha^{ij}|^{\frac{1}{2}n}}{(2\pi)^{\frac{1}{2}np}} \left[\exp - 1/2 \sum_{t=1}^{n} \sum_{i,j=1}^{p} \alpha^{ij} (x_{it} - \mu_i)(x_{jt} - \mu_j)\right] \tag{8.33}$$

where (α^{ij}) is inverse to $(\rho_{ij}\sigma_i\sigma_j)$.

We maximize P for unconstrained values of the parameters to find

$$\hat{\mu}_i = \bar{x}_i \tag{8.34}$$

$$\hat{\alpha}^{ij} = s^{ij} \tag{8.35}$$

and then, on substitution in (8.33) obtain

$$P(\Omega) = \frac{e^{-\frac{1}{2}pn}}{|s_{ij}|^{\frac{1}{2}n} (2\pi)^{\frac{1}{2}pn}} \tag{8.36}$$

We now invert (8.32) to find, on H,

$$\alpha^{ij} = \begin{bmatrix} A & B & \cdots & B \\ B & A & \cdots & B \\ \cdot & \cdot & \cdots & \cdot \\ B & B & \cdots & A \end{bmatrix} \tag{8.37}$$

where

$$A = \frac{1 + (p - 2)\,\rho}{\sigma^2 (1 - \rho)\,\{1 + (p - 1)\,\rho\}} \tag{8.38}$$

$$B = \frac{-\rho}{\sigma^2 (1 - \rho)\,\{1 + (p - 1)\,\rho\}}. \tag{8.39}$$

This gives us, for the likelihood, on substitution in (8.33)

$$P \text{ , say, } = \frac{[(A - B)^{p-1}\,\{A + (p - 1)\,B\}]^{\frac{1}{2}n}}{(2\pi)^{\frac{1}{2}np}} X$$

$$\exp\{-1/2\,[A \Sigma\Sigma (x_{it} - \mu)^2 + B \Sigma\Sigma (x_{it} - \mu)(x_{jt} - \mu)]\}. \tag{8.40}$$

We now find the maximum likelihood estimators

$$\hat{\mu} = \bar{x}\,, \tag{8.41}$$

$$\hat{A} = \frac{1 + (p-2)r_0}{s_0^2\,(1-r_0)\{1 + (p-1)r_0\}}\,, \tag{8.42}$$

$$\hat{B} = \frac{-r_0}{s_0^2\,(1-r_0)\,\{1 + (p-1)r_0\}} \tag{8.43}$$

where

$$s_{ijo} = s_{ij} + (\bar{x}_i - \bar{x})(x_j - \bar{x}), \tag{8.44}$$

$$s_0^2 = 1/p \sum_{i=1}^{p} s_{iio}\,. \tag{8.45}$$

$$r_0 = \frac{\sum_{i \neq j} s_{ijo}}{(p-1)\,\Sigma s_{iio}}\,. \tag{8.46}$$

On substitution in (8.40) we then find

$$p(\omega) = \frac{e^{-\frac{1}{2}pn}}{[(s_0^2)^p\,(1-r_0)^{p-1}\,\{1 + (p-1)r_0\}]^{\frac{1}{2}n}(2\pi)^{\frac{1}{2}pn}} \tag{8.47}$$

We take on our ratio L_{II}

$$\lambda_n^{2/n} = \left(\frac{P(\omega)}{P(\Omega)}\right)^{2/n} = \frac{|s_{ij}|}{(s_0^2)^p\,(1-r_0)^{p-1}\,\{1 + (p-1)r_0\}} \tag{8.48}$$

The denominator is what we get by inserting $r = r_0, s = s_0$ in the matrix of sample dispersions, so that the ratio is, in

fact, a Wilks' ratio of the usual type.

8.7 As before, we require the distribution of the criterion to make a test, and it is arrived at in the usual way. The approximate test states that $-n \log L$ is distributed as χ^2 with $p(p-1)/2 + p + p - 3 = p(p+3)/2 - 3$ degrees of freedom. The corresponding criteria for the other hypotheses are

$$L_1 = \frac{|s_{ij}|}{(s^2)^p (1-r)^{p-1} \{1 + (p-1)r\}} \qquad (8.49)$$

$$L_2 = \frac{s^2(1-r)}{s^2(1-r) + \frac{1}{p-1} \sum_{i=1}^{p} (\bar{x}_i - \bar{x})^2} \qquad (8.50)$$

On the approximate test $-n \log L_1$ is distributed as χ^2 with $\frac{1}{2}p(p+1) - 2$ d.f. and $-n(p-1) \log L_2$ as χ^2 with $p-1$ d.f. The reason for the factor $p-1$ in the latter case is that the criterion L_2 is the $1/2n(p-1)$th root of the λ-ratio.

Example 8.3

In this example (Wilks, 1946) certain details not relevant to the method have been omitted.

A test involving 60 items was given to 100 subjects. The test items were divided into three groups of 20 on the basis of external criteria and a score in each of the three groups obtained for each individual. We may then regard the scores as 100 observations on a trivariate normal population ($p = 3$, $n = 100$). The following values were found :

$\bar{x}_1$	= 10.9900	s_{11}	= 16.8451	s_{12}	= 13.5493
$\bar{x}_2$	= 10.9300	s_{22}	= 18.1099	s_{13}	= 14.5826
$\bar{x}_3$	= 11.2600	s_{33}	= 17.7134	s_{23}	= 13.8056

$$s^2 = 17.5558 \qquad s_o^2 = 17.5764$$

$$r = 0.7963 \qquad r_o = 0.7943$$

$$s_{ij} = 545.5308.$$

The data look homogeneous, that is to say, it appears that the three sub-tests of the original test of 60 items are "parallel". We proceed to examine this hypothesis.

Consider hypothesis H. The criterion of (8.48) becomes 0.9209. The 5% point of $-100 \log L$ for 3 d.f. is 12.5912, giving for the 5% point of L itself 0.8817. The observed value is considerably greater than this and we accept the hypothesis.

Had it been otherwise we might have gone on to test H_1. The criterion of (8.49) becomes 0.9370. The 5% point of $-100 \log L_1$ for 4 d.f. is 9.4877 giving a value of 0.9095 as the 5% point of the criterion itself. The observed value is not significant.

Finally, for the criterion (8.50) we find 0.9914. A test is hardly necessary but if we carry it out the result is not significant.

We conclude that the three sub-tests are similar in respect of means and dispersions.

The analysis of dispersion

8.8 In ordinary variance analysis, tests based on the F-ratio presuppose that error variances are equal. This corresponds to the hypothesis which we have called H_2, and that hypothesis is accordingly the natural extension to multivariate analysis of the univariate analysis of variance. We must here sound a warning about a point which has already occurred in Example 8.2. In variance analysis we often assume equality of underlying variance, sometimes without realizing it, and disaster is apparently averted by the fact that the tests are not very sensitive to small departures from equality. A parallel

assumption that p-variate dispersions are equal is more hazardous. The point has never, so far as I know, been decisively investigated, but the indications are that tests may be somewhat sensitive to differences between parent covariances, though not perhaps to differences between parent variances. To leap at once to hypothesis H_2 before testing H or H_1 is dangerous unless we have prior knowledge about the parental dispersions.

8.9 The analysis of quadratic sums on the basis of hypothesis H_2 is sometimes known as the "multivariate analysis of variance". This is a rather clumsy expression which it may be better to avoid. I shall call it the analysis of dispersion and following Rao (1948a) shall draw a useful distinction between the analysis of *dispersion* and the analysis of *covariance*. In the former we are concerned with the effect of classification on a set of interdependent variates, as illustrated in Examples 8.1 to 8.3. In the latter we also have the effect of classification on a set of variates but we are interested primarily in the effect on one of them. The others appear as disturbing influences and are removed from the principal variate by regression techniques. The distinction is much the same as was drawn earlier between interdependence and dependence. In the general case we may have a complex of $p + g$ variates and wish to study the first p after the effect of the other g has been removed. The removal of the concomitant variation of the g variates is the function of covariance analysis. After it has been performed we may carry out a dispersion analysis on the adjusted p variates, this reducing to a variance analysis if $p = 1$.

8.10 Consider now hypotheses of type H_2 for the p-variate population. If we have k classes the analysis of dispersion may be put in the form

	d.f.	Matrix
Between classes	$k - 1$	(s_{ijm}) say
Within classes	$n - k$	(s_{ija}) say
Total	$n - 1$	(s_{ijo}) say. (8.51)

An analysis on the lines of the foregoing leads to a criterion

λ_n of which the $2/n$th power is

$$W = \frac{|s_{ija}|}{|s_{ijo}|} \tag{8.52}$$

as in (8.21). This may be tested by taking $-n \log W$ as χ^2 with $p(k-1)$ degrees of freedom. In fact, on H_2 the dispersions are common to both Ω and ω; for the former there are pk means and for the latter p means; the number of d.f. is thus $pk - p$.

A similar situation arises if, instead of estimating means we estimate linear functions of them. This result concerning the so-called linear hypothesis enables us to apply (8.51) and the associated test whenever there is a matrix of g $(= k-1)$ degrees of freedom split off from the total matrix and independent of the residual.

8.11 A refinement in the significance test has been proposed by Bartlett, namely that

$$-\{\nu - \tfrac{1}{2}(p+k)\} \tag{8.53}$$

should be taken as χ^2 with $p(k-1)$ d.f. Here ν is the number of degrees of freedom in the total dispersion, our $n-1$.

8.12 For certain small values of p and k these are exact tests in the F-distribution. They are as follows

	Variance ratio		d.f.
$k = 2$, any p	$\frac{1-W}{W}$	$\frac{n-p-1}{p}$	p and $n-p-1$
$k = 3$, any p	$\frac{1-\sqrt{W}}{\sqrt{W}}$	$\frac{n-p-2}{p}$	$2p$ and $2(n-p-2)$
$p = 1$, any k	$\frac{1-W}{W}$	$\frac{n-k}{k-1}$	$k-1$ and $n-k$
$p = 2$, any k	$\frac{1-\sqrt{W}}{\sqrt{W}}$	$\frac{n-k-1}{k-1}$	$2(k-1)$ and $2(n-k-1)$

Box (1949, *Biometrika*, 36, 317) has reviewed and re-examined the distribution theory of criteria of the likelihood-ratio and has obtained better approximations in certain cases. Box's paper contains a number of references to earlier work.

8.13 More complicated situations involving manifold classification and regression can be dealt with by the same technique, provided that we can conduct the analysis so as to emerge with two dispersion determinants independently distributed in Wishart's form; the ratio of one to their sum then follows the W distribution.

Example 8.4.

(Bartlett, 1947 a)

In an experiment to examine the effect of fertilizers on grain 8 treatments were applied in each of 8 randomized blocks; and on each plot two observations were made, the yield of straw (x_1) and the yield of grain (x_2). The following was obtained;

	d.f.	S.S.(x_1^2)	S.P.(x_1x_2)	S.S.(x_2^2)
Blocks	7	86,045.8	56,073.6	75,841.5
Treatments	7	12,496.8	−6,786.6	32,985.0
Residual	49	136,972.6	58,549.0	71,496.1
Total	63	235,515.2	107,836.0	180,322.6

We are not interested in block differences and extract them from the variation. This gives us the middle two lines of the table and a new total:

Total (excluding blocks)	56	149,469.4	51,762.4	104,481.1

Thus

$$W = \frac{\begin{vmatrix} 136,972.6 & 58,549.0 \\ 58,549.0 & 71,496.1 \end{vmatrix}}{\begin{vmatrix} 149,469.4 & 51,762.4 \\ 51,762.4 & 104,481.1 \end{vmatrix}}$$

$$= 0.4920$$

The quantity $-\{56 - \frac{1}{2}(8 + 2)\} \log W = -51 \log 0.4920 = 36.2$ is distributed approximately as χ^2 with $2 \times 7 = 14$ d.f. It is significant, being about on the 0.1% level.

We can proceed to further analysis in several ways. First, in Example 8.1 we can examine the significance of the variates separately by an ordinary F-ratio. This gives

	Estimated Mean Squares		d.f.
	x_1	x_2	
Treatments	1785	4712	7
Residual	2795	1459	49

The first ratio 1.565 for 49 and 7 d.f. is fortunately not significant. (We should have been hard put to it to explain a significance if it had appeared, for the residual mean square is larger than the treatment mean square). The second ratio is 3.230 for 7 and 49 d.f. which is significant at the 1% point. At first sight it would seem that the treatments are affecting grain yield but not straw yield. The correlation "between treatments" is about −0.3 and that "between residuals" is about +0.6, but the former is not significant.

We could also perform a covariance analysis to see if the treatment on straw was influenced by (and perhaps masked by the concomitant variate grain). Our table is

	d.f.	S.S.(x_1^2)	S.P.$(x_1 x_2)$	S.S.(x_2^2)
Treatment	7	12,496.8	−6,786.6	32,985.0
Residual	49	136,972.6	58,549.0	71,496.1
Total	56	149,469.4	51,762.4	104,481.1

Considering the regression of x_1 on x_2 we have for the coefficient in the regression equation (calculated from the residual items)

$$b, \text{ say, } = \frac{S(x_1 x_2)}{S(x_2^2)} = 0.818,911,8$$

The residual total S.S.(x_1^2) is then

$$136,972.6 - 58,549.0\, b = 89.026,1$$

Similarly for the "total" sums of squares and products

$$b', \text{ say } = \frac{51,762.4}{104,481.1} = 0.495,423,6$$

and the residual sum of squares is

$$149,469.4 - 51,762.4\, b' = 128,825.1$$

We then construct the table for residual x_1 effects after the extraction of x_2, obtaining the treatment line by subtraction:

	d.f.	S. S.	Mean Square
Treatment	7	34,799.0	4971
Residual	48	89,026.1	1855
Total	55	123,825.1	2251

We really require to test the ratio of the residual to the total but as this is a univariate test we can equally well test 4971 / 1855 = 2.68 in the F-distribution with 7 and

48 d.f. This is significant at the 5% point. We should infer that the data cannot be wholly explained as an effect on x_2, the grain yield. There appears also an effect on the straw yield which is obscured if we consider that yield by itself, owing to the correlations between the variables.

It may be as well to explain the basis of this covariance analysis. If the jth observation on the ith treatment is y_{ij} our model is

$$y_{ij} = \tau_i + \beta X_{ij}$$

where, for simplicity, y relates to the x_1 variable and X to the x_2 variable. Our hypothesis is not that $\beta = 0$ but that each $\tau_i = 0$, namely that the y's are homogeneous apart from the concomitant X's.

On the hypothesis that $\tau_i = 0$ the estimate of β is given by $S(yX) / S(X^2)$ over the whole sample. This we called b' in our present example and the corresponding residual $S(y^2) - b'S(yX)$ we may call R_T.

If τ_i is not zero, the equations of estimate are

$$\sum_j y_{ij} - \sum_j t_i - b \sum_j (X_{ij}) = 0$$

where t_i is the estimator of τ_i, giving

$$t_i = y_{i.} - b X_{i.},$$

where $y_{i.}$ is the mean of y_{ij}. A second equation of estimate is

$$\sum_{i,j} (yX) - \sum_{i,j} (t_i X_{ij}) - b\, S(X^2) = 0$$

and the residual is

$$S(y^2) - S(t_i\, y_{ij}) - b\, S(y\, X)$$

$$= S(y^2) - S(y_{ij})(y_{i.} - b X_{i.}) - b\, S(y\, X)$$

$$= S(y_{ij})(y_{ij} - y_{i.}) - b\, S\, y_{ij}\, (X_{ij} - X_{i.})$$

$$= S(y_{ij} - y_{i.})^2 - b\, S\, (y_{ij} - y_{i.})(X_{ij} - X_{i.})$$

since

$$\sum_{i,j} (y_{i.})(y_{ij} - y_{i.}) = \sum_{i} y_{i.}(0) = 0$$

This residual, which we may call R_A, is the "within-class" sum as calculated in the example. It is the ratio R_A / R_T which we have, in effect, tested.

Example 8.5

The case of the Egyptian skulls. (Barnard, 1935, *Ann. Eugen. Lond.*, 6, *352*; Kendall, Advanced Theory, vol.2, Chapter 17; Bartlett, 1947a; Rao, 1948b.)

Some data by Miss Barnard have been examined and re-examined by several writers on multivariate analysis and we shall discuss them again here.

Barnard had four series of skulls, 91 from late Predynastic, 162 from the Sixth-to-Twelfth, 70 from the Twelfth- and - Thirteenth and 75 from Ptolemaic dynasties. On each four measurements were taken: x_1 = maximum breadth, x_2 = basial-veolar length, x_3 = nasal height, x_4 = basibregmatic height. The main problems were (a) to construct a discriminant function and (b) to discuss the possible changes in the skull-conformation over time.

(As an additional source of confusion, note that Kendall in earlier editions misdescribed the measurements, Bartlett did so but corrected his 1947a paper before final printing, and Rao, 1948b and in his book, has the variates in the wrong order with a misprint in the mean of x_2 in group II).

The means of the series as given by Miss Barnard are

	I $n_1 = 91$	II $n_2 = 162$	III $n_3 = 70$	IV $n_4 = 75$
x_1	133.582,418	134.265,432	134.371,429	135.306,664
x_2	98.307,692	96.462,963	95.857,143	95.040,000
x_3	50.835,165	51.148,148	50.100,000	52.093,333
x_4	133.000,000	134.882,716	133.642,857	131.466,667

She gives the following sums of squares and products within series (394 d.f.). The column numbers refer to variates.

	1	2	3	4
1	9661.997,470	445.573,301	1130.623,900	2148.584,219
2		9073.115,207	1239.221,990	2255.812,722
3			3938.320,351	1271.054,662
4				8741.508,829

(8.54)

The matrix for the whole observations (397 d.f.) given by Bartlett (1947), and obtained by adding back to (8.54) the matrix between series is:

	1	2	3	4
1	9785.178,098	214.197,666	1217.929,248	2019.820,216
2		9559.460,890	1131.716,372	2381.126,040
3			4088.731,856	1133.473,898
4				9382.242,720

(8.55)

Finally, the matrix between classes (3.d.f.) is

	1	2	3	4
1	123.180,628	-231.375,635	87.305,348	-128.763,994
2		486.345,863	-107.505,618	125.313,318
3			100.411,505	-137.580,764
4				640.733,891

(8.56)

It is as well to look over these results before proceeding.

We note from (8.56) that x_1 and x_3 are positively correlated between classes, as are x_2 and x_4; but that either member of the first pair is negatively correlated with any member of the second pair.

We may, first of all, consider whether there are any significant differences between series on an over-all test. The appropriate criterion is the ratio of the determinants of (8.54) and (8.55), namely

$$L = \frac{2426.898}{2954.474} = 0.8214$$

$-393 \log L = 77.3$ and the number of degrees of freecom is $3 \times 4 = 12$. This value is highly significant and the differences between series we may therefore take as real.

We might now ask whether x_3 and x_4 contribute to this difference independently of the concomitant variation due to x_1 and x_2. This involves extracting the regressions of x_3 and x_4 on x_1 and x_2 from the former and testing the residual matrices.

If y and x are vectors related by

$$y = ax + \varepsilon$$

the regression a is estimated by

$$E(yx') \{E(xx')\}^{-1}$$

and the variation due to regression is

$$394\, E(yx') \{E(xx')\}^{-1} \{E(yx')\}'$$

In our present case, from (8.54)

$$E(xx') = \frac{1}{394} \begin{bmatrix} 9661.997,040 & 445.573,301 \\ 445.573,301 & 9073.115,207 \end{bmatrix}$$

the inverse of which is

$$394 \times 10^{-4} \times \begin{bmatrix} 1.037,332 & -0.050,942 \\ -0.050,942 & 1.104,659 \end{bmatrix}$$

The variation due to regression (factors in 394 cancelling) is then

$$10^{-4} \begin{bmatrix} 1130.623,900 & 1239.221,990 \\ 2148.584,210 & 2255.812,722 \end{bmatrix} \begin{bmatrix} 1.037,332 & -0.050,942 \\ -0.050,942 & 1.104,659 \end{bmatrix} \times \begin{bmatrix} 1130.623,900 & 2148.584,210 \\ 1239.221,990 & 2255.812,722 \end{bmatrix}$$

$$= \begin{bmatrix} 287.967,620 & 534.238,796 \\ 534.238,796 & 991.621,041 \end{bmatrix} \qquad (8.57)$$

Subtracting this from the matrix of x_3 and x_4, viz

$$\begin{bmatrix} 3938.320,351 & 1271.054,662 \\ 1271.054,662 & 8741.508,829 \end{bmatrix}$$

we get the residual

$$\begin{bmatrix} 3650.353,731 & 736.815,866 \\ 736.815,866 & 7749.887,788 \end{bmatrix}$$

with $394-2 = 392$ d.f. (8.58)

Similarly, operating on (8.55) for the total dispersions we find the residual

$$\begin{bmatrix} 3809.335,190 & 611.698,381 \\ 611.698,381 & 8393.755,848 \end{bmatrix} \qquad (8.59)$$

with $397-2 = 395$ d.f.

The ratio of the determinants of (8.58) to (8.59) is

$$L = \frac{0.277,469}{0.316,003} = 0.8781$$

$-392 \log L = 51.39$ and the appropriate number of degrees of freedom is $3 \times 2 = 6$. The value is highly significant and we conclude that x_3 and x_4 are "relevant" variates, in that the differences between series cannot be assigned to x_1 and x_2 alone. Another way of looking at this will emerge when we consider discriminant functions.

One of the topics discussed by Miss Barnard was the use of these measurements in discriminating between the four periods and the possibility of the variates having a linear regression on time (that is to say a linear trend which might, perhaps, be different for the four variates). The intervals between the four series were taken in the proportion 2, 1, 2 and we may conveniently take the values of t as -5, -1, $+1$, $+5$. On this basis the sum of products of x_1, x_2, x_3, x_4 with time $t - \bar{t}$ are

718.762,86, −1407.260,75, −410.101,94, −733.427,58 and

$\Sigma(t - \bar{t})^2 = 4307.663,32$. ($\bar{t}$ is not zero because the numbers in classes are unequal, and its value is −0.432,161)

We are now examining, not the regression of some variates on others, but the regression of all on the extraneous variable time. The sums of squares and products due to regression (1 d.f.) are

	1	2	3	4
1	119.930,358	−234.810,812	68.428,235	−122.377,258
2		459.734,449	−133.975,163	−149.601,596
3			39.042,852	−69.824,358
4				124.874,099

(8.60)

Here, for example, the item in row 1 and column 1 is

$$\frac{\{\Sigma x_1 (t - \bar{t})\}^2}{\Sigma(t - \bar{t})^2} = \frac{718.762,86^2}{4307.668,32} = 119.930,358$$

and that in row 1 and column 2 is

$$\frac{\Sigma x_1(t-\bar{t})\ \Sigma x_2(t-\bar{t})}{\Sigma(t-\bar{t})^2} = \frac{718.762,86 \times -1407.260,75}{4307.668,32} = -234.810,812$$

The residual after removing the regression on time from the original matrix is given by subtracting (8.60) from (8.55), giving (396 d.f.)

	1	2	3	4
1	9665.247,740	449.008,478	1149.501,013	2142.197,474
2		9099.726,441	1265.691,535	2231.524,444
3			4049.689,004	1203.298,256
4				9257.368,621

(8.61)

We now consider whether this residual is homogeneous, testing the variation within series as given by (8.54) - 394 d.f. against (8.61). The criterion is

$$L = \frac{10^{12} \times 0.242,691}{10^{12} \times 0.268,738} = 0.9031$$

The multiplier of $-\log L$ is $396 - 1/2(2+4+1) = 392\ 1/2$, and the d.f. number $2 \times 4 = 8$. $-392\ 1/2 \log .9031 = 40.02$ which is significant.

We conclude that either (a) the regression on time is not linear or (b) that there are additional sources of variation between series which are not temporal effects.

8.14 In conclusion reference may be made to the work of Roy on the comparison of two dispersion matrices of the same order p. If the matrices are S_1 and S_2 the roots of

$$| S_1 - \lambda S_2 | = 0 \qquad (8.62)$$

which, for non-degenerate cases, are those of

$$| S_1 S_2^{-1} - \lambda I | = 0 \qquad (8.63)$$

will have roots near unity if they emanated from the same populations. Roy proposes to reject the hypothesis of equality if, when the roots are arranged in *ascending* order

$$0 \leqslant \lambda_1 \leqslant \lambda_2 \quad \cdots \leqslant \lambda_p \leqslant \infty . \qquad (8.64)$$

$\lambda_p \geqslant \lambda_0$ and / or $\lambda_1 \leqslant \lambda_0'$, where λ_0 and λ_0' are given by

$$P(\lambda_p \geqslant \lambda_0) = P(\lambda_1 \leqslant \lambda_0') \text{ and} \qquad (8.65)$$

$$1 - P(\lambda_0 \leqslant \lambda_1 \leqslant \lambda_p \leqslant \lambda_0') = \text{say, } 0.05, \qquad (8.66)$$

the probabilities being calculated on the null hypothesis.

The same type of test may be applied in dispersion analysis. If the "mean-square" matrices are S_1 and S_2, based on $k - 1$ and $n - k$ degrees of freedom, S_1 will be positive semi-definite and of rank q, say, where q is the smaller of p and $k - 1$; while S_2 is positive definite. The roots of (8.63) will have $p - q$ zeros and q non-zeros, and the latter may be tested in the usual way. Although great progress has been made with the distribution problem tables only began to become available in 1957. (See Foster, F.G., *Biometrika*, 45.)

9. DISCRIMINATORY ANALYSIS

9.1 Suppose that we have two p-variate populations of a similar kind which "overlap" in the sense that certain members can be observed which might have arisen from either population. There will be occasions when, confronted with a member and noting its variate-values, we are uncertain from which population it emanated. We will suppose that we have to make up our minds on this question and require to allocate it to one or the other of the parents. By what rule should we proceed so as to make as few mistakes as possible over a large number of similar occasions ? Questions of this type give rise to *discriminatory analysis*, the general object of which is to find rules of behaviour in the assignment of individuals to predetermined classes with optimal properties.

Note two things: (a) the classes are predetermined. It is not the object of the inquiry to find what is the best way of dividing heterogeneous material into populations or classes; these are already decided.

(b) We shall, in general, consider the assignment of a member to one population or the other: extensions of the theory which leave room for suspended judgement will not be considered, though they are quite possible.

9.2 We shall consider in the first instance the case where there are only two parents; and we shall seek for some functions of the variates (linear for simplicity if possible) which will assume widely different sets of values for the two, so that a sample observation can be allotted to the appropriate population according to the value which it gives to the function. We shall then consider the more extended problems

(a) when there are k parents and (b) when we allow ourselves several discriminant functions, not merely one.

9.3 When we allot a member to one of two populations, A_1 and A_2, we may make two kinds of mistakes, according to the population to which we wrongly assign a given member. We shall suppose that these two kinds of mistakes are equally important. (In the contrary case we should have to weight the errors proportionately to their importance, as in the general theory of decision functions). We shall also assume that we require the two kinds of mistake to occur with equal, proportional frequency for each population, and that their sum should be a minimum.

Consider a space of p dimensions in which a sample member is represented by a point. The populations may be imagined as clusters of points (or densities in the continuous case) condensing round two central values. We wish to set up a boundary, or rather a region R in the space such that, if $f_1\,dx$, $f_2\,dx$, represent the elements of frequency of A_1 and A_2

$$\int_R f_2\,dx = \int_{1-R} f_1\,dx$$

$$= 1 - \int_R f_1\,dx \qquad (9.1)$$

This is equivalent to

$$\int_R (f_1 + f_2)\,dx = 1 \qquad (9.2)$$

We require, subject to this condition, to minimize

$$\int_R f_2\,dx \qquad (9.3)$$

This is equivalent to finding an unconditional minimum of

$$\int_R \{f_2 - \lambda\,(f_1 + f_2)\}\,dx$$

or equivalently again, of

$$\int (\beta f_2 - f_1)\, dx$$

where the constant is determined by (9.2). This is clearly achieved by taking into R all those points for which $\beta f_2 - f_1$ is negative. Thus the boundary of R is given by

$$\frac{f_1}{f_2} = \beta \tag{9.4}$$

9.4 From this point of view the discriminating boundary arises naturally as determined by the probability ratio. Any point for which $f_1/f_2 > \beta$ is assigned to f_1, and in the contrary case to f_2. The probabilities of misclassification either way are then equal and minimal; and each is equal to

$$\int_{f_1/f_2 > \beta} f_2\, dx = \int_{f_2/f_1 > \beta} f_1\, dx. \tag{9.5}$$

Formally, at least, this solves the simplest version of our problem.

9.5 Now suppose that the two populations are multivariate normal with means μ_{1i}, μ_{2i} and identical dispersion matrices α_{ij}. The logarithm of the likelihood ratio, apart from constants, is

$$-\frac{1}{2} \sum_{i,j=1}^{p} \alpha^{ij}\{(x_i - \mu_{1i})(x_j - \mu_{1j}) - (x_i - \mu_{2i})(x_j - \mu_{2j})\}$$

$$= \Sigma\{\alpha^{ij}\,(\mu_{1j} - \mu_{2j})\, x_i - \tfrac{1}{2}\, \alpha^{ij}(\mu_{1i}\mu_{1j} - \mu_{2i}\mu_{2j})\}$$

The second part of this expression is a constant and without losing generality we can take as our function

$$\Sigma \alpha^{ij}\,(\mu_{1j} - \mu_{2j})\, x_i. \tag{9.6}$$

This is the parental form. In practice we usually do not know the parameters but have to estimate them from a sample

for which the parents are known. Inserting the sample values in (9.6) – they are maximum-likelihood estimators of the corresponding parameters – we have for the discriminant function X

$$X = \Sigma\, a^{ij}\, (\bar{x}_{1j} - \bar{x}_{2j})\, x_i. \qquad (9.7)$$

9.6 The same result may be reached by a different route. Suppose that we determine a linear function

$$X = \sum_{j=1}^{p} l_j\, x_j$$

so as to maximize the square of the difference of its expectation in the two populations, divided by the variance of X (which, by hypothesis, is common to the two populations). That is to say, we maximize

$$\frac{\{\sum_j l_j\, (\mu_{1j} - \mu_{2j})\}^2}{\sum_{i,j} l_i\, l_j\, \alpha_{ij}}$$

A differentiation with respect to l_j gives us

$$(\mu_{1j} - \mu_{2j}) = \{\sum_{j=1}^{p} l_j\, (\mu_{1j} - \mu_{2j})\}\, \Sigma\, l_i\, \alpha_{ij} / 2 \Sigma \alpha_{ij}\, l_i\, l_j$$

from which we have

$$l_i \propto \Sigma\, a^{ij}\, (\mu_{1j} - \mu_{2j})$$

leading back to (9.6). Since our function X is used only to separate the two populations, not to measure the distance between them, we may multiply it by any convenient constant.

Example 9.1

(Fisher, 1936)

Measurements were made on 50 flowers from each of two species of iris, *setosa* and *versicolor*. Four measurements were made, sepal length, sepal width, petal length and petal

width. We denote them by the suffixes 1, 2, 3, 4.

The sums of squares and products about the means were (in $cm.^2$)

	1	2	3	4
1	19.1434	9.0356	9.7634	3.2394
2		11.8658	4.6232	2.4746
3			12.2978	3.8794
4				2.4604

(9.8)

The means were

	Versicolor	Setosa	V − S
1	5.936	5.006	0.930
2	2.770	3.428	−0.658
3	4.260	1.462	2.798
4	1.326	0.246	1.080

(9.9)

The matrix inverse to (9.8) is ($cm.^{-2}$)

	1	2	3	4
1	.188,716,1	−.006,866,6	−.081,615,8	.039,635,0
2		.145,273,6	.033,410,1	.110,752,9
3			.219,361,4	−.272,020,6
4				.894,550,6

(9.9a)

Using (9.7) we then find, for the coefficients

$$l_1 = (.118,716,1)(0.930) - (.066,866,6)(-.658) - (.081,615,8)(2.798) + (.039,635,0)(1.080)$$

$$= -0.031,151,1$$

$$l_2 = -0.183,907,5$$

$$l_3 = 0.222,104,4$$

$$l_4 = 0.314,737,4$$

We may conveniently choose multiples of these coefficients so that the coefficient of x_1 is unity. We then find

$$X = x_1 + 5.9037x_2 - 7.1299x_3 - 10.1036x_4 \qquad (9.10)$$

The mean value of X for versicolor, obtained by substituting from (9.9) in this, is -21.4815; and the mean for setosa is 12.3345. The difference is 33.8160 cm.

From (9.6) we have, relating X to the original constants l

$$X = \Sigma\, \alpha^{ij}\, (\mu_{1j} - \mu_{2j})x_i = \Sigma\, l_i\, x_i$$

Then

$$\begin{aligned} \text{var } X &= \Sigma\, l_i\, l_j \text{ cov } (x_i, x_j) \\ &= \Sigma\, l_i\, l_j\, \alpha_{ij} \\ &= \sum_{i,j} l_i\, \alpha_{ij} \sum_k \alpha^{jk}(\mu_{1k} - \mu_{2k}) \\ &= \Sigma\, l_i\, (\mu_{1i} - \mu_{2i}) \\ &= \bar{X}_1 - \bar{X}_2. \end{aligned}$$

We may then estimate var X from the difference of the observed mean values. In our present case the variance of X of (9.10) is given by

$$\frac{33.8160}{(0.031,151,1)n} = 1085.54/n$$

where n is the number of degrees of freedom of the estimate

and is 98. Thus var X = 11.08. This is the variance of a single value. The variance of the difference of two means, each of 50 numbers, is 1/25 th of this, namely 0.4432, giving a standard error of 0.664. The observed difference of means, 33.816, is more than 50 times as great as the standard error and we conclude that the discriminant is likely to be effective.

The probability of misclassification is easily estimated. It is the integral of the multivariate form over a region to one side of the plane X = constant, which is easily seen to be the tail of a normal distribution function of X. The standard deviation of X is $\sqrt{11.08}$ = 3.32 and the distance between the parent means in the X – direction is 33.816. One half of this is 16.9 and we require the tail area of a normal error in excess of 16.9 / 3.38 standard deviations from the mean. This is negligible.

Example 9.2

(F. Heincke, 1898, Naturgeschichte des Herings, Berlin; S. R. Zarapkin, 1934, *Arch. Naturgesch.* **3**, 161; G. Beall, 1945, *Psychometrika*, **10**, 205; L.S.Penrose, 1947)

Biometricians have often proposed to discriminate between individuals on the basis of "size" and "shape". Consider the case where measurements are made on an organism and the correlations between them are positive and all equal, say, to r. The correlation matrix is then the one that we considered in Example 2.2 and has latent roots $\lambda_1 = 1 + (p-1)\,r$, $\lambda_2 = \ldots = \lambda_p = 1 - r$. From the component analysis approach this means that there is one principal direction and that the others are isotropic. The component corresponding to λ_1 .

$$\zeta_1 = \frac{1}{\sqrt{p}} \Sigma x_i . \tag{9.11}$$

We take a "size" component proportional to this and write

$$Q = \Sigma x_i = \bar{\zeta}_1 \sqrt{p} \tag{9.12}$$

so that

$$\text{var } Q = p\lambda_1 = p\{1 + (p-1)r\}. \qquad (9.13)$$

Among the remaining variation no particular direction is suggested as suitable. Let us take a set of weights w_i with non-zero mean w and define a shape component by

$$P = \sum_{j=1}^{p} \frac{w_j - w}{w} x_j. \qquad (9.14)$$

We have at once

$$\text{var } P = (1-r) \Sigma \left(\frac{w_j - w}{w}\right)^2 \qquad (9.15)$$

where the x's are taken to have standard measure.

Also

$$\begin{aligned} \text{cov } (Q,P) &= \text{cov } \{ \Sigma \frac{w_j - w}{w} x_j, \Sigma x_j \} \\ &= \sum_j \frac{w_j - w}{w} \text{var } x_j + \sum_{i \neq j} \frac{w_j - w}{w} \text{cov } (x_i, x_j) \\ &= \{ 1 + (p-1) r \} \Sigma \frac{w_j - w}{w} \\ &= 0 \end{aligned} \qquad (9.16)$$

The shape component is then uncorrelated with the size component and this will remain approximately true if the correlations are nearly equal but not exactly so.

When we are interested in discrimination we may choose the weights so as best to discriminate between two populations and hence arrive at an *ad hoc* measure of "shape". We shall take

$$w_j = \bar{x}_{1j} - \bar{x}_{2j} \qquad (9.17)$$

and shall look for a linear function of size and shape which is the best discriminator :

$$X = \alpha Q + P \tag{9.18}$$

This will be given by determining α so as to maximize

$$\frac{(X_1 - X_2)^2}{\text{var } X}$$

If $D_P = P_1 - P_2$ and $D_Q = Q_1 - Q_2$, the suffixes as usual referring to the two populations, this requires the maximization of

$$\frac{(\alpha D_Q + D_P)^2}{\alpha^2 \text{ var } Q + 2\alpha \text{ cov } (Q, P) + \text{var } P}$$

leading to

$$\alpha = \frac{D_Q \text{ var } P - D_P \text{ cov } (Q, P)}{D_P \text{ var } Q - D_Q \text{ cov } (Q, P)}. \tag{9.19}$$

But Q and P are uncorrelated. We have also

$$D_P = \sum_{j=1}^{p} \frac{\bar{x}_{1j} - \bar{x}_{2j} - (\bar{x}_1 - \bar{x}_2)}{\bar{x}_1 - \bar{x}_2} (x_{1j} - x_{2j})$$

$$= p \left\{ \sum \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{\bar{x}_1 - \bar{x}_2} - (\bar{x}_1 - \bar{x}_2) \right\},$$

$$\text{var } P = (1 - r) \sum \left(\frac{x_{1j} - x_{2j} - x_1 - x_2}{\bar{x}_1 - \bar{x}_2} \right)^2$$

$$= (1 - r) \left\{ \sum \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{(\bar{x}_1 - \bar{x}_2)^2} - (\bar{x}_1 - \bar{x}_2) \right\},$$

$$D_Q = p\,(\bar{x}_1 - \bar{x}_2),$$

$$\text{var}\,Q = p\,\{\,1 + (p-1)\,r\,\}.$$

On substitution for α in (9.19) we then find

$$X = \frac{1-r}{1+(p-1)\,r}\,Q + P \qquad (9.20)$$

This, in the sense we have defined, is the "best" linear function of size and shape.

Reverting to the Iris data of Example 9.1 we have the following values of the mean when expressed in terms of common standard deviation and reduced to common zero means:

	Versicolor	Setosa	V − S
1	1.0628	−1.0628	2.1256
2	−0.9551	0.9551	−1.9102
3	3.9894	−3.9894	7.9788
4	3.4426	−3.4426	6.8852
Sum = Q	7.5397	−7.5397	D_Q = 15.0794
Var Q	10.5076	10.5076	(9.21)

The estimate of size, Q, is simply the sum of the standardized variates for each variety. The variance of Q is calculated from the dispersion matrix reduced by standardization to a correlation matrix :

	1	2	3	4
1	1.	.599,513	.636,323	.472,011
2		1.	.382,719	.457,988
3			1.	.705,258
4				1. (9.22)

Thus var Q is the sum of the 16 elements in this matrix.

The weightings for shape are found from (9.21) by dividing the figures in the last column by $\frac{1}{4}$ (15.0794) and subtracting unity, and are −0.4362, −1.5067, 1.1165, 0.8264. For the estimate of shape we then find

$$P = (-0.4362 \times 1.0628) + \text{etc.}$$

$$= 8.2747.$$

and by using (9.22) again we find

$$\text{var } P = 3.0912.$$

The covariance of Q and P does not exactly vanish and we calculate it from (9.22) as 0.36162. Thus, substituting in (9.19) we find for α

$$\alpha = \frac{(15.0794)(3.0912) - (16.5494)(0.36162)}{(16.5494)(10.5076) - (15.0794)(0.36162)}$$

$$= 0.2412$$

Hence the discriminator is

$$X = 0.2412\,Q + P. \qquad (9.23)$$

Example 9.3

(C. A. B. Smith, 1947)

The method of the foregoing example is useful in reducing the discriminant to a function of two variables only; and if we work on size and shape variables P and Q we can handle quadratic discriminators without too much mathematical complexity.

A group of 25 normal and 25 psychotic individuals were given certain tests, and for each individual a size and shape variable x and y were determined. The results were—

25 Normals			25 Psychotics		
x	y	z	x	y	z
22	6	62	24	38	8
20	14	36	19	36	−13
23	9	61	11	43	−67
23	1	77	6	60	−126
17	8	33	9	32	−55
24	9	66	10	17	−52
23	13	53	3	17	−55
18	18	28	15	56	−73
22	16	42	14	43	−52
19	18	23	20	8	48
20	17	30	8	46	−88
20	31	2	20	62	−60
21	9	51	14	36	−38
13	13	3	3	12	−45
20	14	36	10	51	−88
19	15	29	22	22	30
20	11	42	11	30	−41
18	17	20	6	30	−66
20	7	50	20	61	−58
23	6	67	20	43	−22
23	23	33	15	43	−57
25	4	71	5	53	−107
23	5	69	10	43	−72
21	12	43	13	19	−9
23	7	65	12	4	26
Totals 520	308	1084	320	910	−1120
Mean 20.8	12.32	43.4	12.80	36.40	−44.8
					(9.24)

Here z is a quantity $5x - 2y - 36$ to be explained later. We have the following quantities:

	Normal	Psychotics
Mean of x	20.80	12.80
Mean of y	12.32	36.40
Var x	6.92	36.75
Var y	40.89	287.92
Cov (x,y)	-5.27	13.92
d-f	24	24
S. d of x	2.63	6.06
S. d of y	6.39	16.97
Correlation	-0.31	0.14

(9.25)

Let us consider first of all a linear discriminant function of x and y. Pooling the dispersions we get for the supposed dispersion matrix

$$\begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} 21.83 & 4.33 \\ & 164.40 \end{bmatrix} \end{array} \qquad (9.26)$$

and for the difference of means (Normal - Psychotics):

x	8.00
y	-24.08.

The inverse of (9.26) is proportional to

$$\begin{bmatrix} 164.40 & -4.33 \\ -4.33 & 21.83 \end{bmatrix}$$

The discriminant, from (9.7), is then

$$(164.40 \times 8.00 + 4.33 \times 24.08)(x_1 - 20.8)$$
$$+ (-4.33 \times 8.00 + 21.83 \times -24.08)(x_2 - 12.80)$$
$$= 1419x - 560y - 10,198$$

on division by 280 this becomes, nearly enough,

$$z = 5x - 2y - 36. \qquad (9.27)$$

The values of this function are shown in (9.24). It is seen that z is positive for all normals (no errors) and negative for all psychotics except four (16% error). The errors of classification, as estimated from the data themselves, are accordingly not symmetrical and amount to 8% over all. This is better than we should do by using x or y alone; for instance, if we take $x \geqslant 17$ to be normal and $x \leqslant 16$ to be psychotic there would be 1 error in classifying the normals and 6 for the psychotics.

We may remark, however, that the variances and covariances of x and y are very different for the two types; and we doubt whether it is legitimate to suppose that they have a common dispersion matrix. If we go back to the approach of (9.5), but assume different dispersion matrices α and β, say, the logarithm of the likelihood ratio becomes proportional to

$$\Sigma\alpha^{ij}\,(x_i - \mu_{1i})(x_j - \mu_{1j}) - \Sigma\beta^{ij}\,(x_i - \mu_{2i})(x_j - \mu_{2j}) \qquad (9.28)$$

which is a quadratic function. We now use the fact that size and shape have a correlation which can be put equal to zero. The expression (9.28) then reduces to a form of type

$$(\xi - k_1)^2 + (\eta - k_2)^2 \qquad (9.29)$$

where

$$\xi = x\sqrt{(\alpha^{11} - \beta^{11})} \text{ etc.}$$

In our present case we have for the estimates of α, β

$$a^{11} = \frac{1}{6.92} = 0.1445$$

$$a^{22} = 0.0244$$

$$b^{11} = 0.0272$$

$$b^{22} = 0.0035$$

and the discriminant function becomes

$$-0.1173\ (x - 22.65)^2 - 0.0209\ (y - 8.29)^2 + 3.16. \quad (9.30)$$

On multiplication by 2 this becomes, nearly enough

$$X = \left(\frac{x - 23}{2}\right)^2 + \left(\frac{y - 8}{5}\right)^2 - 16 \quad (9.31)$$

The values given to the 25 normals by this function are −16, −12, −16, −14, −7, −16, −15, −6, −13, −8, −10, 7, −15, 10, −12, −10, −13, −7, −14, −16, −7, −15, −16, −14, −16. (2 positive, error = 8%).

The values given to the psychotics are 20, 19, 69, 164, 56, 29, 87, 92, 53, −14, 95, 103, 36, 85, 100, −8, 39, 76, 99, 41, 64, 146, 75, 14, 5. (2 negative, error = 8%).

It is instructive to plot the data (as Smith does) on a diagram and to examine how the points lie in relation to the discriminatory lines.

The significance of a discriminant function

9.7 We may ask whether a discriminator is "significant". Such a question needs a little clarification. We may mean that there is a real difference between the populations but that they are so close together that a discriminator is not very effective; this is measured by the errors of

misclassification which, though minimal, may still be large. Or we may mean that there is a real difference between the populations but our sample size is not large enough to produce a very reliable discriminator; this is really a matter of setting confidence intervals to the function or its coefficients. Or we may mean that the parents are identical and that a discriminant function is illusory.

9.8 Questions of "significance" in discriminant functions have usually been discussed in terms of the last possibility. They are not so much tests of the functions as tests of homogeneity by the use of the functions. If heterogeneity is found the function, *ipso facto*, is significant in the sense that it discriminates between real differences in an optimal way (except that we use estimators of dispersions and means instead of the unknown parent values). But that way may not be very good even if it is the best.

9.9 Suppose our two populations have, in fact, identical means. The difference of the means in the discriminator is then

$$U, \text{ say, } = \Sigma a^{ij} (\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1i} - \bar{x}_{2i}) \qquad (9.32)$$

The term $\bar{x}_{1j} - \bar{x}_{2j}$ is the difference of two means, each normally distributed, and is therefore distributed like a mean about zero with twice the variance of a single mean if $n_1 = n_2$. It follows that U is distributed as Hotelling's $T^2/(2n - 1)$, based on $2n$ observations. This is equivalent to the distribution of multiple R in the null case and can be carried out by an analysis of variance.

9.10 The same conclusion is reached from the following approach. We now generalize slightly to the case where there are n_1 members observed in one class and n_2 in the other. It is readily verified from 9.5 that this does not affect the form (9.7) for the discriminant function. We introduce a dummy for a *dependent* variate y on a pseudo-regression equation

$$y = \Sigma l_j (x_j - \bar{x}_j) \qquad (9.33)$$

by putting

$$y = \frac{n_2}{n_1 + n_2}, \; -\frac{n_1}{n_1 + n_2} \tag{9.34}$$

according as the member falls in the first or second class. The mean of y is zero and in (9.33) we take $\bar{x}_j$ as the mean of x_j over both classes, so that

$$\bar{x}_j = \frac{n_1 \bar{x}_{1j} + n_2 \bar{x}_{2j}}{n_1 + n_2} \tag{9.35}$$

Treating (9.35) as a regression we find

$$S\{y\,(x_j - \bar{x}_j)\} = \sum_i \{l_i S\,(x_i - \bar{x}_i)(x_j - \bar{x}_j)\}$$

where summation S refers to sample values.

Now from (9.34)

$$S\{y\,(x_j - \bar{x}_j)\} = \frac{n_2}{n_1 + n_2} n_1\,(\bar{x}_{1j} - \bar{x}_j) - \frac{n_1}{n_1 + n_2} n_2 (\bar{x}_{2j} - \bar{x}_j)$$

$$= \frac{n_1 n_2}{n_1 + n_2}\,(\bar{x}_{1j} - \bar{x}_{2j}). \tag{9.36}$$

We also have

$$\begin{aligned} S\,(x_j - \bar{x}_j)(x_i - \bar{x}_i) = {} & S_1\,(x_i - \bar{x}_{1i})(x_j - \bar{x}_{1j}) \\ & + S_2\,(x_i - \bar{x}_{2i})(x_j - \bar{x}_{2j}) \\ & + S_1\,(\bar{x}_{1i} - \bar{x}_i)(\bar{x}_{1j} - \bar{x}_j) \\ & + S_2\,(\bar{x}_{2i} - \bar{x}_i)(\bar{x}_{2j} - \bar{x}_j) \end{aligned}$$

where S_1, S_2 refer to summation over the first and second groups; and by use of (9.35) this is reduced to

$$= (n_1 + n_2)\,a_{ij} + \frac{n_1 n_2}{n_1 + n_2}\,(\bar{x}_{1i} - \bar{x}_{2i})(\bar{x}_{1j} - \bar{x}_{2j}). \tag{9.37}$$

Hence we find, putting

$$K = \Sigma l_i \, (\bar{x}_{1i} - \bar{x}_{2i}), \tag{9.37a}$$

$$\frac{n_1 \, n_2}{n_1 + n_2} \, (1 - K) \, (\bar{x}_{1i} - \bar{x}_{2i}) = \Sigma \, (n_1 + n_2) \, a_{ij} \, l_i,$$

giving

$$l_i = \frac{n_1 \, n_2}{(n_1 + n_2)^2} \, (1 - K) \, \Sigma a^{ij} \, (\bar{x}_{1j} - \bar{x}_{2j}) \tag{9.38}$$

leading once again to the discriminant function. The constant K is given by putting these values in (9.37a) as

$$K = \frac{n_1 \, n_2}{(n_1 + n_2)^2} (1 - K) \, \Sigma a^{ij} \, (\bar{x}_{1i} - \bar{x}_{2i})(\bar{x}_{1j} - \bar{x}_{2j}). \tag{9.39}$$

We also find

$$S \, (y^2) = \frac{n_1 \, n_2}{n_1 + n_2} \tag{9.40}$$

The regression analysis may then be written

Sum of Squares	d.f.
$\frac{n_1 \, n_2}{n_1 + n_2}$ times	
$\Sigma l_i (\bar{x}_{1i} - \bar{x}_{2i})$	p
$1 - \Sigma l_i (\bar{x}_{1i} - \bar{x}_{2i})$	$n_1 + n_2 - p - 1$
1	$n_1 + n_2 - 1$

(9.41)

Now it does <u>not</u> follow immediately from ordinary regression theory that this can be tested as a variance analysis. We

have put dummy variates for our dependent variable and the ordinary theory applies to models where the dependent variable is a random variate and the independent variables may be fixed (and in particular, dummies). It is, however, a remarkable fact that the test still applies as a test of homogeneity. The basic reason is a kind of duality which exists in the sample space. In ordinary regression theory for the null case (no parental multiple correlation) we find the distribution of the angle between a random (dependent) vector and a fixed (independent) plane. On the present occasion we required the distribution of the angle between the fixed (dependent) vector and the random (independent) plane. And the two amount to the same thing. When we leave the null case this duality, in general, breaks down.

Example 9.4

Let us revert to the Iris data of Example 9.1 and 9.2. We found values of l's in the former, but in applying (9.41) we have to be careful about ceofficients. Those l's were found from the deviance (not the dispersion matrix). Using them and (9.7) we find

$$\Sigma l_i\,(\bar{x}_{1i} - \bar{x}_{2i}) = (-0.031,151,1)(0.930) + (-0.183,907,5)(-0.658) + (0.222,104,4)(2.758) + (0.314,737,0)(1.080)$$

$$= 1.053,404,683,2.$$

Since $n_1 = n_2 = 50$ we have for K, from (9.39)

$$K = \frac{1.053,404,68}{1.053,404,68 + 1/25}$$

$$= 0.963,417$$

Thus the analysis is

	S.S. 25 times	d.f.	Quotient
"Regression"	0.963,417	4	0.240,854
Residual	0.036,573	95	0.000,385
Total	1.000,000	99	

(9.42)

and the "regression" is overwhelmingly significant.

We may, if we wish, test the significance of individual coefficients in the discriminant functions by regarding them as regression coefficients. We use the matrix inverse to (9.8), namely (9.9a). For example the value of l_i in Example 9.1 is −0.031,151,1. We standardize it by multiplying by $\frac{n_1\, n_2}{n_1+n_2}(1-K)$, namely, 0.914,325, to obtain −0.028,482. In (9.9) the term corresponding to l_1 is 0.118,716,1, which must be multiplied by the residual quotient in (9.42), namely .000,384,979 × 25 to give an estimate of the variance of the coefficient as .001,142,580 with a standard error of .0336. The actual value is rather less than this and we are led to doubt whether x_1 is playing any important part in the discrimination.

We may remark that in this example, although the value of the discriminant function would be slightly impaired if we discarded any variate, except perhaps x_1, we can get a discriminant which is quite good enough for ordinary purposes by using x_3, petal length, alone. The variance of x_3, from (9.8), is estimated as 12.2978/98 = 0.125,488. The mean difference (of two sets of 50) has then a variance of .005,019,52 with a standard error of .0785. On this scale the discriminant function would be x_3 itself. The mean difference in two sets of 50 is then 2.798, about 36 times the standard error; not so big as the factor of 50 for the discriminant function based on four variates but big enough for practical purposes. The error of misclassification is about equal to the tail area of a normal curve to the right of an ordinate four standard deviations to the right of the mean.

The case of k populations

9.11 When we proceed from discrimination between two populations to discrimination among a member of populations an essentially new point appears. As before, we shall endeavour to divide up the sample space into mutually exclusive regions, one for for each population, and allot an observed member to the population in whose region it falls. But the boundaries of the

regions are no longer determined by one single discriminant function. Either we must, to achieve optimal properties, have several functions, or, if we must have a single function, we shall have to sacrifice some discriminatory power.

9.12 It will be enough for expository purposes if we consider three populations – the generalization to k is immediate. We will also generalize to the extent of supposing that the probabilities of occurrence of the three populations whose density functions are f_1, f_2, f_3 are respectively π_1, π_2, π_3 ($\pi_1 + \pi_2 + \pi_3 = 1$). If the corresponding regions are R_1, R_2, R_3, a generalization by Rao of a lemma due in essence to Neyman and Pearson states that the errors of misclassification are a minimum if the regions are determined by probability ratios which form a simple extension of 9.3. In fact R_1 is such that $\pi_1 f_1$ is greater than or equal to both $\pi_2 f_2$ and $\pi_3 f_3$; R_2 is such that $\pi_2 f_2 \geqslant \pi_3 f_3$ and $\pi_1 f_1$; R_3 is such that $\pi_3 f_3 \geqslant \pi_1 f_1$ and $\pi_2 f_2$.

9.13 In particular, if the three populations are normal with common dispersion matrix α_{ij} and means μ_{1i}, μ_{2i}, μ_{3i}, it follows as in the manner of 9.5 that R_1 must be such that

$$\Sigma\alpha^{ij}\,(\mu_{1j} - \mu_{2j})\,x_i \geqslant \text{some constant } \beta_{12}, \text{ say;} \qquad (9.43)$$

$$\Sigma\alpha^{ij}\,(\mu_{ij} - \mu_{3j})\,x_i \geqslant \text{some constant } \beta_{13}, \text{ say.} \qquad (9.44)$$

Similarly for the other regions. In the sample R_1 will be determined as the domain lying between the two hyper-planes (9.43) and (9.44) and including the mean of population 1; and so on. The surfaces of constant weighted probability ratio for populations 1 and 2 are, in fact, given by

$$\log \frac{\pi_1 f_1}{\pi_2 f_2} = \Sigma\alpha^{ij}\,(\mu_{1j} - \mu_{2j})x_i - \tfrac{1}{2}\Sigma\alpha^{ij}\,(\mu_{1i}\mu_{2j} - \mu_{2i}\mu_{2j}) + \log \pi_1/\pi_2 \qquad (9.45)$$

In the particular case where all the π's are equal we may compare the three functions

$$X_1 = \Sigma\alpha^{ij}\,\mu_{1j}\,x_i - \tfrac{1}{2}\,\Sigma\alpha^{ij}\,\mu_{1i}\,\mu_{1j} \qquad (9.46)$$

$$X_2 = \Sigma\alpha^{ij}\,\mu_{2j}\,x_i - \tfrac{1}{2}\,\Sigma\alpha^{ij}\,\mu_{2i}\,\mu_{2j} \qquad (9.47)$$

$$X_3 = \Sigma\alpha^{ij}\,\mu_{3j}\,x_i - \tfrac{1}{2}\,\Sigma\alpha^{ij}\,\mu_{3i}\,\mu_{3j} \qquad (9.48)$$

and allot a member to R_1, R_2, R_3 according to which of the Xs is the greatest when the sample values are substituted. For if, say X_1 is the greatest, it follows from (9.45) that $f_1 > f_2$ and $f_1 > f_3$. As usual we may substitute sample values for the unknown parameters in these equations to get an approximate discriminator.

Example 9.5

(Rao and Slater, 1949, *Brit. Jour. Psych.* **2**, 17)

A number of persons falling into certain neurotic groups obtained the following mean scores in three tests x_1, x_2, x_3.

Group	Sample Size	Mean Score 1	2	3
Anxiety state	114	2.9298	1.1667	0.7281
Hysteria	33	3.0303	1.2424	0.5455
Psychopathy	32	3.8125	1.8438	0.8125
Obsession	17	4.7059	1.5882	1.1176
Personality change	5	1.4000	0.2000	0.0000
Normal	55	0.6000	0.1455	0.2182
	256			(9.49)

The *dispersion* matrix within groups (250 d.f.) was

	1	2	3	
1	2.300,851	0.251,578	0.474,169	
2		0.607,466	0.035,774	
3			0.595,094	(9.50)

Its inverse is

	1	2	3
1	0.543,234	−0.200,195	−0.420,813
2		1.725,807	0.055,767
3			2.012,357

(9.51)

For the purposes of this example I will suppose all the π's to be equal. The functions of type (9.46) then are as follows :

	Coefficients			
	x_1	x_2	x_3	Constant
Normal	0.2050	0.1431	0.1947	−0.0931
Personality change	0.7204	0.0649	−0.5780	−0.5107
Anxiety state	1.0515	1.4676	0.2974	−2.5047
Hysteria	1.1678	1.5679	−0.1081	−2.7139
Psychopathy	1.3599	2.4641	0.1336	−4.9182
Obsession	1.7680	1.8611	0.3573	−5.8375

(9.52)

Here, for example, the coefficient of x_1 for the normal state is

$$(0.543,234)(0.6000) - (0.200,195)(0.1455) + (-0.420,813)(0.2182)$$
$$= 0.2050.$$

Suppose, for example, we had a subject with scores 1, 1, 0. The values of the functions, in the order of (9.52) are 0.2550, 0.2746, 0.0144, .0218, −1,0942, −2.2084. We assign the member to the second group, personality change. In practice, of course, we should do so very tentatively. The normal group is very close and there are only five members in the personality-change group on which the sample discriminators are based.

9.14 Consider again the geometrical representation of the

situation in a p-way space. The k populations, assumed with identical dispersions, are centered at k points in the space and we have been discussing the partitioning of that space into k parts.

Now if the population-means all lie on a straight line and the π's are equal all the functions of type (9.46) are proportional and we can use any one as a discriminator. We can then partition the space by parallel planes; or, looked at from a slightly different view-point, can measure our discriminating function along the line of means. In short, we reduce the problem to one involving only a single discriminator.

9.15 In practice it will happen only rarely that this ideal situation arises; but as an approximation we might find the line of closest fit to the means and use it as one discriminator; and if this is not sufficient, find a second orthogonal line as a second discriminator; and so on. This, in point of fact, leads us to a method very akin to component analysis.

Let us suppose that we have k populations of p variates, and require to determine the constants in a function

$$X = \Sigma l_j x_j$$

such that the ratio of variances between and within classes is maximized. This is, in fact, the procedure we have already followed for $k = 2$ in (9.6). It comes to the same thing, and is more convenient, to maximize the ratio of 'between' to 'total' variance. If (A) represents the matrix between means and (B) the matrix within classes we shall have to maximize

$$\lambda = \frac{\Sigma A_{ij} l_i l_j}{\Sigma B_{ij} l_i l_j} \qquad (9.53)$$

which leads to

$$\sum_i (A_{ij} - \lambda B_{ij}) l_i = 0 \qquad (9.54)$$

giving the familiar determinantal form

$$| A - \lambda B | = 0 \qquad (9.55)$$

The largest root of this provides our discriminant function. We may, in the manner mentioned below in 9.16, proceed to test whether further roots are required for additional discriminators.

Example 9.6

(M.S. Bartlett, 1951, *Ann. Eugen. Lond.*, **16**, 199; E. J. Williams, 1952, *Biometrika*, **39**,17)

Let us reconsider the data of Example 8.4 concerning yield of straw (x_1) and grain (x_2) after the elimination of block-effects. We found for the matrix A between treatments (7 d.f.)

$$\begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & \quad 2 \end{array} \\ \begin{bmatrix} 12{,}496.8 & -6{,}786.6 \\ & 32{,}985.0 \end{bmatrix} \end{array} \qquad (9.56)$$

and for the totals B (56 d.f.)

$$\begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & \quad 2 \end{array} \\ \begin{bmatrix} 149{,}469.4 & 51{,}762.4 \\ & 104{,}481.1 \end{bmatrix} \end{array} \qquad (9.57)$$

Equation (9.55) then becomes

$$\begin{vmatrix} 12{,}496.8 - 149{,}469.4\,\lambda & -6{,}786.6 - 51{,}762.4\,\lambda \\ -6786.6 - 51{,}762.4\,\lambda & 32{,}985.0 - 104{,}481.1\,\lambda \end{vmatrix} = 0$$

a quadratic in λ with roots (9.58)

$$\lambda_1 = 0.47698, \quad \lambda_2 = 0.05934.$$

The l's corresponding to λ_1 are (proportionally) given by either of

$$58{,}797.1\ l_1 + 31{,}477.2\ l_2 = 0$$

$$31{,}477.2\ l_1 + 16{,}850.4\ l_2 = 0$$

Taking the coefficient of l_2 as unity we find $l_1 = -0.535$ and the discriminant is

$$x_2 - 0.535\ x_1 \qquad (9.59)$$

9.16 This aspect of the subject has been carried further by E. J. Williams (*loc. cit.*, 1952) who derived an exact test of significance of a discriminant function and by M.S.Bartlett (*loc. cit.* 1951) who suggested some approximate tests and extended Williams' results. The investigations are similar to those mentioned earlier in component analysis; the object is to see what is the smallest number of dimensions in which the parent means lie. If they are collinear, one discriminator is sufficient; if they are coplanar two are required; and so on. The general treatment thus links up with component and canonical correlation analysis and our various topics are to be seen as different aspects of the same fundamental structure.

9.17 A few final notes

(a) We remarked in chapter 8 that concomitant variation could be abstracted by a regression technique and a dispersion-analysis carried out on the adjusted variables. The same is true for discriminatory analysis. A worked example is given by Cochran and Bliss (1948). As a general rule, however, there seems little to gain by eliminating 'internal' variation in this way. The eliminated variates might as well be retained in this discriminant if they are to be used at all.

(b) For further discussion of the use of dummy *dependent* variates in constructing analyses of variance see Cochran and Bliss (1948); and for the use of ordinary regression-theory tests on coefficients in a discriminant function see Bartlett (1939a).

(c) For discrimination by the D^2—statistic see Rao's book on *Advanced Statistical Methods in Biometric Research.*

(d) Discrimination problems often arise with variates which are not measurable, e.g. simple dichotomies or classifications. These are often treated by inserting (0,1) variables or (−1,0,1) variables but the method is rather rough.

REFERENCES

Those marked * may conveniently form the starting points of further reading.

Aitken, A.C. (1937) The evaluation of the latent roots and vectors of a matrix. *Proc. Roy. Soc. Ed.* A, 37, 269.

Anderson, T. W. (1946) The non-central Wishart distribution and certain problems of multivariate statistics. *Ann. Math. Statist.*, 17, 409.

Anderson, T. W. (1948) The asymptotic distributions of certain determinantal equations. *J. R. Statist. Soc.* B, 10, 132.

Anderson, T. W. (1951 a). Classification by multivariate analysis. *Psychometrika*, 16, 31.

Anderson, T. W. (1951 b). The asymptotic distribution of certain characteristic vectors. *Proc. Second Berkeley Symposium*, University of California Press.

Anderson, T. W. (1951 c). Estimating linear restrictions on regression coefficient for multivariate normal distributions. *Ann. Math. Statist.*, 22, 327.

Anderson, T. W. and Girschik, M. A. (1944). Some extensions of the Wishart distribution. *Ann. Math. Statist.*, 15, 345.

Anderson, T. W. and Rubin, H. (1949). Estimation of parameters of a single equation. *Ann. Math. Statist.*, 20, 46.

Anderson, T. W. and Rubin, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.*, 21, 570.

*Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings Third Berkeley Symposium*, 5, 111.

Bartlett, M. S. (1934). The vector representation of a sample. *Proc. Camb. Phil. Soc.*, 30, 327.

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Proc. Camb. Phil. Soc.*, 34, 33.

Bartlett, M. S. (1939 a). The standard errors of discriminant function coefficients. *J. Roy. Statist. Soc. Supp.*, 6, 169.

Bartlett, M. S. (1939 b). A note on tests of significance in multivariate analysis. *Proc. Camb. Phil. Soc.*, 35, 180.

Bartlett, M. S. (1941). The statistical significance of canonical correlations. *Biometrika*, 32, 29.

Bartlett, M. S. (1947 a). Multivariate analysis. *J. Roy. Statist. Soc.*, B, 9, 176.

Bartlett, M. S. (1947 b). The general canonical correlation distribution. *Ann. Math. Statist.*, 18, 1.

Bartlett, M. S. (1948). Internal and external factor analysis. *Brit. J. Psych. (Stat. Sect.)*, 1. 73.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *Brit. J. Psych. (Stat. Sect.)*, 3, 77.

Bartlett, M. S. (1951 a). A further note on tests of significance in factor analysis. *Brit. J. Psych. (Stat. Sect.)*, 4, 1.

Bartlett, M. S. (1951 b). The effect of standardization on an approximation in factor analysis. *Biometrika*, 38, 337.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *J. Roy. Statist. Soc.*, B., 16, 296.

Berkson, J. (1950). Are there two regressions ? *J. Am. Stat. Ass.*, 45, 164.

Bhattacharyya, A. (1946). On a measure of divergence of two multinomial populations. *Sankhya*, 7, 401.

Birnbaum, Z. W. and Chapman, D. G. (1950). On optimum selections from multinormal populations. *Ann. Math. Statist.*, 21, 443

Bishop, D. J. (1939). On a comprehensive test of the homogeneity of variances and covariances in multivariate problems. *Biometrika*, 31, 31.

Bose, R. C. (1936). On the exact distribution and moment coefficients of the D^2-statistic. *Sankhya*, 2, 143.

Bose, R. C. and Roy, S. N. (1938). The distribution of the Studentized D^2-Statistic. *Sankhya*, 4, 19.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36, 317.

Brown, G. W. (1947). Discriminant functions. *Ann. Math. Statist.*, 18, 514.

Buckatzsch, E. J. (1947). The influence of social conditions on mortality rates. *Population Studies*, 1, 229.

Burt, Sir Cyril (1937). Method of factor analysis with and without successive approximations. *Brit. J. Ed. Psych.*, 7, 172.

*Burt, Sir Cyril (1949). Alternative methods of factor analysis. *Brit. J. Psych. (Stat. Sect.)*, 2, 98.

Burt, Sir Cyril and Banks, C. H. (1941). A factor analysis of body measurements for British adult males. *Ann. Eugen. Lond.*, 13, 238.

Camp, B. H. (1932). The converse of Spearman's two-factor theorem. *Biometrika*, 24, 418.

Cochran, W. G. (1943). The comparison of different scales of measurements for experimental results. *Ann. Math. Statist.*, 14, 205.

Cochran, W. G. and Bliss, C. I. (1948). Discriminant functions with covariance. *Ann. Math. Statist.*, 19, 151.

Creasy, M. A. (1956). Confidence limits for the gradient in the linear functional relationship. *J. Roy. Statist. Soc.*, B, 18, 65.

Dick, I. D. (1952). The equivalent of factor analysis confluence analysis in problems of multicollinearity. *New Zealand Jour. Sci. Tech.*, B, 33, 245.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen. Lond.*, 7, 179.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Ann. Eugen. Lond.*, 5, 376.

Fisher, R. A. (1939). The sampling distribution of some statistics obtained from non-linear regression. *Ann. Eugen. Lond.*, 9, 238.

Fisher, R. A. (1940). The precision of discriminant functions. *Ann. Eugen. Lond.*, 10, 422.

Geary, R. C. (1948). Studies in relations between economic time-series. *J. Roy. Statist. Soc.*, B, 10, 140.

Girschik, M. A. (1939). On the sampling theory of the roots of determinantal equations. *Ann. Math. Statist.*, 10, 203.

Gosnell, H. F. and Schmidt, Margaret (1936). Factorial and correlational analysis of the 1934 vote in Chicago. *J. Am. Statist. Assoc.*, 31, 507.

Harper, R. et al. (1950). The application of multiple factor analysis to industrial test data. *Brit. Journ. App. Physics*, 1, 1.

Hoel, P. G. (1937). A significance test for component analysis. *Ann. Math. Statist.*, 8, 149.

Hoel, P. G. (1939). A significance test for minimum rank in factor analysis. *Psychometrika*, 4, 245.

*Holzinger, K. J. and Harman, H. H. (1941). *Factor Analysis*. Chicago U.P.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.*, 2, 360.

Hsu, P. L. (1938). Note on Hotelling's generalized T. *Ann. Math. Statist.*, 9, 231.

Hsu, P. L. (1939). On the distributions of the roots of certain determinantal equations. *Ann. Eugen. Lond.*, 9, 250.

Hsu, P. L. (1940). On generalized analysis of variance, I. *Biometrika*, 31, 221.

Hsu, P. L. (1941 a). On the limiting distribution of the canonical correlations. *Biometrika*, 32, 38.

Hsu, P. L. (1941 b). On the limiting distribution of the roots of determinantal equations. *Jour. Lond. Math. Soc.*, 16, 183.

Hsu, P. L. (1941 c). On the problem of rank and the limiting distribution of Fisher's test function. *Ann. Eugen. Lond.*, 11, 39.

Hsu, P. L. (1941 d). Canonical reduction of the general regression problem. *Ann. Eugen. Lond.*, 11, 42.

James, G. S. (1954). Test of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41, 19.

Kelley, T. L. (1935). Essential traits of mental life. *Harvard U.P.*

Kendall, M. G. (1939). The geographical distribution of crop productivity in England. *J. Roy. Stat. Soc.*, 102, 21.

Kendall, M. G. and Babington Smith, B. (1950). Factor analysis. *J. Roy. Statist. Soc.* B, 12, 60.

Lawley, D. N. (1938). A generalization of Fisher's z-test. *Biometrika*, 30, 180.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Ed.*, A, 60, 64.

Lawley, D. N. (1942). Further investigations in factor estimation. *Proc. Roy. Soc. Ed.*, A. 61, 176.

Lawley, D. N. (1947). Problems in factor analysis. *ibid.* A, 62, 394.

Lawley, D. N. (1953). A modified method of estimations in factor analysis and some large sample results. *Uppsala Symposium*, 35.

*Lawley, D. N. (1955). A statistical examination of the centroid method. *Proc. Roy. Soc. Ed.*, A. 64, 175.

Lawley, D. N. (1956 a). Test of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43, 128.

*Lawley, D. N. (1956 b). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43, 295.

Lawley, D. N. and Swanson, Z. (1954). Tests of significance in a factor analysis of artificial data. *Brit. Jour. Stat. Psy.*, 7, 75.

Lederman, W. (1937). On the rank of the reduced correlational matrix in multiple factor analysis. *Psychometrika*, 2, 83.

Lederman, W. (1939). On a shortened version of estimation of mental factors by regression. *Psychometrika*, 4, 109.

Lukomski, J. (1939). On some properties of multidimensional problems. *Ann. Math. Statist.*, 10, 236.

Mahalanobis, P. C. (1948). Historical note on the D^2-statistic. *Sankhya*, 9, 237.

Mahalanobis, P. C., Bose, R. C. and Roy, S. N. (1937). Normalization of statistical variates and the rise of rectangular coordinates in the theory of sampling distributions. *Sankhya*, 3, 35.

Mises, R. von (1945). On the classification of observation data into distinct groups. *Ann. Math. Statist.*, 16, 68.

Mood, A. M. (1951). On the distribution of the characteristic roots of normal second-moment matrices. *Ann. Math. Statist.*, 22, 266.

Nanda, D. N. (1948). Distribution of a root of a determinantal equation. *Ann. Math. Statist.*, 19, 47; and Limiting distribution of a root of a determinantal equation. *Ann. Math. Statist.*, 19, 340.

Pearson, E. S. and Wilks, S. S. (1933). Method of statistical analysis appropriate for k samples of two variables. *Biometrika*, 25, 353.

Penrose, L. S. (1947). Some notes on discrimination. *Ann. Eugen. Lond.*, 13, 228.

Pillai, K. C. S. (1956). On the distribution of the largest or the smallest root of a matrix in multivariate analysis. *Biometrika*, 43, 122.

Plackett, R. L. (1947). An exact test for the equality of variances. *Biometrika*, 34, 311.

Rao, C. R. (1946). Tests with discriminant functions in multivariate analysis. *Sankhya*, 7, 407.

Rao, C. R. (1948 a). The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc.*, B, 10, 159.

Rao, C. R. (1948 b). Tests of significance in multivariate analysis. *Biometrika*, 35, 58.

Rao, C. R. (1948 c). On the distance between two populations. *Sankhya*, 9, 246.

Rao, C. R. (1949). On some problems arising out of discrimination with multiple characters. *Sankhya*, 9, 343.

Rao, C. R. (1950). Statistical inference applied to classificatory problems. *Sankhya*, 10, 229.

*Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research.* New York, John Wiley & Sons.

Rhodes, E. C. (1937). An index of business activity. *J. Roy. Stat. Soc.*, 100, 18.

Roy, S. N. (1939). p-statistics, or some generalizations in analysis of variance appropriate to multivariate problems. *Sankhya*, 4, 381.

Roy, S. N. (1941). Analysis of variance for multiple populations: the sampling distribution of the requisite p-statistics on the null and non-null hypothesis. *Sankhya*, 6, 35.

Roy, S. N. (1943). The individual sampling distribution of the maximum, the minimum and any intermediate of the 'p' -statistics on the null hypothesis. *Sankhya*, 7, 133.

Roy, S. N. (1948) Notes on the testing of composite hypotheses – II. *Sankya*, 9,19.

Roy, S. N. (1950). Univariate and multivariate analysis as problems in the testing of composite hypotheses – I. *Sankhya*, 10, 29.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.*, 24, 220.

Slater, P. (1947). The factor analysis of a matrix of 2 x 2 tables. *J. Roy. Statist. Soc. Supp.*, 9, 1.

Smith, C. A. B. (1947). Some examples of discrimination. *Ann. Eugen. Lond.*, 13, 272.

Stone, J. R. N. (1947) The interdependence of blocks of transactions. *J. Roy. Stat. Soc. Supp.*, 9, 1.

Thomson, Sir Godfrey (1939). *The Factorial Analysis of Human Ability*. London U.P.

*Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago U.P.

Tintner, G. (1945). A note on rank, multicollinearity and multiple regression. *Ann. Math. Stats.*, 16, 304.

Uppsala Symposium on Psychological Factor Analysis (1953), Stockholm, Almquist and Wiksell.

Vernon, P. E. (1949). How many factors ? Private manus.

Wald, A. and Brookner, R. J. (1941). On the distribution of Wilks' statistic for testing independence of several groups of variables. *Ann. Math. Statist.*, 12, 137.

Waugh, F. W. (1942). Regression between sets of variates. *Econometrica*, 10, 290.

Welch, B. L. (1939). Note on discriminant functions. *Biometrika*, 31, 218.

Whittle, P. (1953). On principal components and least square methods of factor analysis. *Skand. Akt.*, 35, 223.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24, 471.

Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, 3, 309.

Wilks, S. S. (1946). Sample criteria for testing equality of means, etc. in a normal multivariate system. *Ann. Math. Statist.*, 17, 257.

Wilks, S. S. and Tukey, J. W. (1946). Approximation of the distribution of the product of beta variables by a single beta variable. *Ann. Math. Statist.*, 17, 318.

Young, Gale (1941). Maximum likelihood estimation and factor analysis. *Psychometrika*, 6, 49.

EXERCISES

1. A p-variate complex has the following correlation matrix:

$$\begin{bmatrix} 1 & r & r^2 & \dots & r^{p-1} \\ r & 1 & r & \dots & r^{p-2} \\ r^2 & r & 1 & \dots & r^{p-3} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ r^{p-1} & r^{p-2} & r^{p-3} & \dots & 1 \end{bmatrix}$$

Show that the determinant of the matrix is $(1 - r^2)^{p-1}$ and hence that, apart from the trivial case when $r = 1$, the complex cannot be represented in fewer than p dimensions.

2. The correlations of a variate x_1 with x_2 and x_3 respectively are 0.8 and 0.6. Find the correlation between x_2 and x_3 if the variation of the three variates can be expressed in two dimensions.

3. A p-variate complex has the correlation matrix

$$\begin{bmatrix} 1 & r & r & \dots & r \\ r & 1 & r & \dots & r \\ \cdot & \cdot & \cdot & \dots & \cdot \\ r & r & r & \dots & 1 \end{bmatrix}$$

Show that if $r > 0$ it has one greatest characteristic root and that all the others are equal. Verify that the sum of all the roots is p.

4. The three variables (Sweden, 1921 – 1938)

x_1 = log price of lump sugar

x_2 = log animal food price

x_3 = log income per head

have the correlation matrix

$$\begin{bmatrix} 1.0 & .631555 & -.867190 \\ & 1.0 & -.330909 \\ & & 1.0 \end{bmatrix}$$

Show that the three characteristic roots of this matrix are 2.245,548, .688,417 and .066,036.

5. Do a centroid analysis on the matrix

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$$

by reflecting x_1, x_2 and x_3 at the second, third and fourth stages respectively and hence derive the factors.

$$f_1 = \sqrt{\frac{5}{8}}\,(x_1 + x_2 + x_3 + x_4)$$

$$f_2 = \frac{1}{\sqrt{24}}\,(-3x_1 + x_2 + x_3 + x_4)$$

$$f_3 = \frac{1}{\sqrt{12}}\,(\quad -2x_2 + x_3 + x_4)$$

$$f_4 = \frac{1}{2}\,(\qquad\quad -x_3 + x_4)$$

6. In question 4, regard the expression $|r - \lambda I| = f(\lambda)$ as a cubic in λ and evaluate it for four convenient values of λ, e.g. 0, 1, 2, 3. Hence find $f(\lambda)$ and solve it to obtain the characteristic roots.

7. Do a centroid analysis on the correlation matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

by reflecting x_1, x_2 and x_3 respectively at successive stages, and hence derive the transformation to new variables y which are also uncorrelated and have unit variance :

$$y_1 = \frac{1}{2}(x_1 + x_2 + x_3 + x_4)$$

$$y_2 = \frac{1}{\sqrt{12}}(-3x_1 + x_2 + x_3 + x_4)$$

$$y_3 = \frac{1}{\sqrt{6}}(-2x_2 + x_3 + x_4)$$

$$y_4 = \frac{1}{\sqrt{2}}(-x_3 + x_4)$$

(This is a particular case of a well-known transformation known as Helmert's.)

8. A set of observations (x, y) are subject to independent errors with equal variances. If the estimate of a linear relation between them based on maximum likelihood is $y = x \tan \psi$ and the two regression lines which are obtained by regarding one or the other variable as fixed are $y = x \tan \theta_1$, $x = y \tan \theta_2$, show that

$$2 \cot 2\psi = \cot \theta_1 - \tan \theta_2$$

9. In the previous exercise, show that the maximum likelihood estimate of the "true" value of a pair (x_i, y_i) is obtained by dropping a perpendicular from that point on to the estimated line of linear relationship.

10. Verify by writing the equations explicitly that (4.39) is insufficient for the estimation of the β's when the distributions are normal.

11. A pair of variates y_1, y_2 are uncorrelated; a set of three variates x_1, x_2, x_3 are also uncorrelated. Each of the y's is correlated with each of the x's to an equal extent, the correlation coefficients being r. By finding the greatest canonical correlation show that r cannot exceed $1/\sqrt{6}$ in absolute value.

12. In the problem of k samples from bivariate normal populations, show that in the usual notation the criterion for testing the hypothesis H_2 (that the k populations, given equal dispersions, have equal means) is

$$\lambda_{H_2} = \left\{ \frac{|\, v_{ija} \,|}{|\, v_{ijo} \,|} \right\}^{\frac{1}{2}n}$$

where n is the total sample number.

Show how to test the significance of this criterion.

13. Four strains of mice are treated with a drug and, after a certain period, on each mouse the following variates are measured in suitable units:

x_1 = gain in weight
x_2 = iron content of blood

The sum of squares and cross product about means are as follows :

	Strains			
	1	2	3	4
Mean x_1	2	3	4	3
$S(x_1 - \bar{x}_1)^2$	17	16	28	12
Mean x_2	15	18	16	20
$S(x_2 - \bar{x}_2)^2$	56	36	60	45
$S(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$	24	10	30	10
Sample number	10	8	14	11

Perform an analysis of dispersion to test the differences of means of the populations.

14. In the previous exercise use covariance analysis to discuss the homogenity of the strains (a) according to x_1 after removal of the effect of x_2; (b) according to x_2 after the removal of x_1.

Comment on your results.

15. Explain what is implied in the expression "testing the significance of components" in component analysis.

Carry out the process on the characteristic roots of Example 4 and state what you suppose your results to mean. (n=18).

16. Taking the first two variates of Example 9.1, show that a discriminant function is

$$x_1 - 1.236\, x_2$$

and that this is practically as good a discriminator as the four-variate discriminant function of the example.

17. It is desired to discriminate between individuals drawn from two populations, Π_1 and Π_2, on the basis of measurement of some, or all, of the characters x_1, x_2, x_3. The costs of measuring x_1, x_2 and x_3 on an individual are respectively 2s., 4s., and 7s.

The variance - covariance matrix of x_1, x_2 and x_3, which is the same for both Π_1 and Π_2, is

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{bmatrix} 4 & & \\ 3 & 9 & \\ 1 & 0 & 1 \end{bmatrix}$$

and the expected values of x_1, x_2 and x_3 in Π_1, Π_2 respectively are

	x_1	x_2	x_3
Π_1	5	10	8
Π_2	0	2	4

The loss involved in the misclassification of an individual may be assessed at 10s. Find the most economical discriminant function to use if the populations occur with equal frequency. It may be assumed that the joint distribution of x_1, x_2, x_3 is multinormal.

(London B.Sc. Special, 1956)